

Kelli D. Cummings  
Yaacov Petscher *Editors*

# The Fluency Construct

Curriculum-Based Measurement  
Concepts and Applications

 Springer

# The Fluency Construct

Kelli D. Cummings • Yaacov Petscher  
Editors

# The Fluency Construct

Curriculum-Based Measurement  
Concepts and Applications

 Springer

*Editors*

Kelli D. Cummings  
University of Maryland  
College Park  
Maryland  
USA

Yaacov Petscher  
Florida Center for Reading Research  
Florida State University  
Tallahassee  
Florida  
USA

ISBN 978-1-4939-2802-6

ISBN 978-1-4939-2803-3 (eBook)

DOI 10.1007/978-1-4939-2803-3

Library of Congress Control Number: 2015943073

Springer New York Heidelberg Dordrecht London

© Springer Science+Business Media, LLC 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media  
([www.springer.com](http://www.springer.com))



*For my brother Kyle and my parents Gary and Diane who have always supported my love for books and school. I would also like to thank my academic advisor, Dr. Kenneth W. Merrell, and my very first graduate school professor, Dr. Mark Shinn, who taught me everything I know about problem solving and CBM.*

Kelli D. Cummings, Ph.D., NCSP

*For my nephew Jacob and my nieces Madelyn, Lucy, Nora, Lilah, and Gwen.*

Yaacov Petscher, Ph.D.

# Preface

This book represents the collective work of approximately 18 research groups actively engaged in fluency-based curriculum-based measurement (CBM) work across the country and internationally. Its release coincides with two recent journal special issues on the topic of fluency-based measurement technology (Cummings & Biancarosa, 2015; Petscher, Cummings, Biancarosa, & Fien, 2013) and represents, at least a portion of, work from second- and third-generation research labs investigating the development, implementation, and interpretation of fluency-based measurement technology in schools. Though initiated as a special education technology, CBM has been both directly and indirectly part of many fundamental paradigm shifts in education since its practice was first codified during the mid-1970s through the University of Minnesota Institute for Research on Learning Disabilities (IRLD; Deno & Mirkin, 1977). Notable contributions include: (a) the shift in school psychology practice from a within-student, aptitude-by-treatment interaction (ATI) approach to one of formative assessment within the context of more effective instruction (Deno, 1990); (b) the move toward universal screening of all general education students to make early intervention more powerful and effective (NCLB, 2001; Reading First); (c) reauthorizing the Individuals with Disabilities Education Act (2004; Gersten et al., 2008); and (d) continuing the push toward data-based decision-making regarding student progress and efficacious programs through the American Recovery and Reinvestment Act (ARRA, 2009). For many of us, whether we loved or hated these new uses, they have left an indelible print on our landscape that is hard to ignore. As we look toward the future, with the Common Core State Standards (CCSS) for instruction, next generation assessments, computer-adaptive testing, and ever-growing issues with the measurement technology, some may ask what the future holds for CBM. We hope to answer some of those questions here.

The purpose of this book is to provide a comprehensive overview of fluency as construct applied through the use of CBM technology. Biancarosa and Shanley provide an introductory foundation to the text in Chap. 1 by introducing us to the concept and definitions of fluency from the perspective of three educational areas (i.e., language acquisition, reading, and mathematics), and closing with recommendations for improved clarity regarding the term “fluency” across fields. The book is then organized into sections based on the primary interest group that is targeted

by the content. In Part 1, we focus on educational professionals who use fluency-based measurement data to make decisions about their students. In Part 2, we turn our focus to the area of test development, with chapters focused on test equating and methods used to select criterion-referenced benchmark goals. As an additional focus in this section, we bring to bear important work in the areas of classical test and item-response theories, which hold critical implications for CBM test construction in the future. Part 3 of the text deals with advanced statistical methods for measurement researchers utilizing fluency data. The text as a whole is closed by Espin and Deno Chap. 13, who remind us that fluency measures (as well as educational assessment in general) must remain grounded in both the decisions we wish to make as well as the consequences, both intentional and unintentional, of those decisions. By targeting the diverse groups of fluency CBM users and researchers, we hope to paint a picture of the construct that is nuanced and relevant for the myriad decisions that fluency data are intended to facilitate. Nevertheless, as with most scientific endeavors, the true value of this book is the foundation it will provide to future work. We hope that this text, at least a small part of it, sets the stage for your own participation in the future research and development of fluency CBMs.

# Contents

<b>1 What Is Fluency?</b> .....	1
Gina Biancarosa and Lina Shanley	
<b>Part I Applied Use of Fluency Measures</b>	
<b>2 Indicators of Fluent Writing in Beginning Writers</b> .....	21
Kristen D. Ritchey, Kristen L. McMaster, Stephanie Al Otaiba, Cynthia S. Puranik, Young-Suk Grace Kim, David C. Parker and Miriam Ortiz	
<b>3 Mathematics Fluency—More than the Weekly Timed Test</b> .....	67
Ben Clarke, Nancy Nelson and Lina Shanley	
<b>4 Using Curriculum-Based Measurement Fluency Data for Initial Screening Decisions</b> .....	91
Erica S. Lembke, Abigail Carlisle and Apryl Poch	
<b>5 Using Oral Reading Fluency to Evaluate Response to Intervention and to Identify Students not Making Sufficient Progress</b> .....	123
Matthew K. Burns, Benjamin Silbergliitt, Theodore J. Christ, Kimberly A. Gibbons and Melissa Coolong-Chaffin	
<b>Part II Recommendations for Test Developers</b>	
<b>6 Foundations of Fluency-Based Assessments in Behavioral and Psychometric Paradigms</b> .....	143
Theodore J. Christ, Ethan R. Van Norman and Peter M. Nelson	
<b>7 Using Response Time and Accuracy Data to Inform the Measurement of Fluency</b> .....	165
John J. Prindle, Alison M. Mitchell and Yaacov Petscher	

**8 An Introduction to the Statistical Evaluation of Fluency Measures with Signal Detection Theory**..... 187  
Keith Smolkowski, Kelli D. Cummings and Lisa Strycker

**9 Different Approaches to Equating Oral Reading Fluency Passages ...** 223  
Kristi L. Santi, Christopher Barr, Shiva Khalaf and David J. Francis

**Part III Advanced Research Methods**

**10 Using Individual Growth Curves to Model Reading Fluency** ..... 269  
D. Betsy McCoach and Huihui Yu

**11 Introduction to Latent Class Analysis for Reading Fluency Research** 309  
Jessica A. R. Logan and Jill M. Pentimonti

**12 Using Latent Change Score Analysis to Model Co-Development in Fluency Skills**..... 333  
Yaacov Petscher, Sharon Koon and Sarah Herrera

**13 Conclusion: Oral Reading Fluency or Reading Aloud from Text: An Analysis Through a Unified View of Construct Validity**..... 365  
Christine A. Espin and Stanley L. Deno

**Index**..... 385

## About the Editors

**Kelli D. Cummings, Ph.D., NCSP**, is an assistant professor at the University of Maryland. Her research focuses on projects that link assessment and intervention technologies to improve student success. She has worked as a special education teacher, a school psychologist, and a trainer of school psychologists; she brings these practical experiences to her work in the research community. She has provided formal technical assistance and training on problem solving, response-to-intervention, and the use of curriculum-based measurement for data-based decision-making in schools throughout the USA, Canada, and Great Britain. Her work is published regularly in school psychology and education journals.

**Yaacov Petscher, Ph.D.**, is the director of research at the Florida Center for Reading Research at Florida State University. His research interests are in the areas of computer adaptive assessments, the study of individual differences in reading, psychometrics, and applied research methods in education. He recently edited *Applied Quantitative Analysis in Education and the Social Sciences*, and his works are regularly published in education, psychology, and measurement journals as well as technical reports for the National Center for Education Evaluation and Regional Assistance via the Regional Educational Laboratory-Southeast.

# Contributors

**Stephanie Al Otaiba** Southern Methodist University, Dallas, TX, USA

**Christopher Barr** Texas Institute for Measurement, Evaluation, and Statistics,  
University of Houston, Houston, TX, USA

**Gina Biancarosa** Center on Teaching and Learning, University of Oregon, Eugene,  
OR, USA

**Matthew K. Burns** University of Missouri, Columbia, MO, USA

**Abigail Carlisle** University of Missouri, Columbia, MO, USA

**Theodore J. Christ** University of Minnesota, Minneapolis, MN, USA

**Ben Clarke** Center on Teaching and Learning, University of Oregon, Eugene, TX,  
USA

**Melissa Coolong-Chaffin** University of Wisconsin Eau Claire, Eau Claire, WI,  
USA

**Kelli D. Cummings** University of Maryland, College Park, USA

**Stanley L. Deno** University of Minnesota, Minneapolis, MN, USA

**Christine A. Espin** Leiden University, Leiden, South Holland, The Netherlands

**David J. Francis** Texas Institute for Measurement, Evaluation, and Statistics,  
University of Houston, Houston, TX, USA

**Kimberly A. Gibbons** St. Croix River Education District, Rush City, MN, USA

**Sarah Herrera** Florida Center for Reading Research, Florida State University,  
Tallahassee, FL, USA

**Shiva Khalaf** College of Education, University of Houston, Houston, TX, USA

**Young-Suk Grace Kim** Florida State University, Tallahassee, FL, USA

**Sharon Koon** Florida Center for Reading Research, Florida State University, Tallahassee, FL, USA

**Erica S. Lembke** University of Missouri, Columbia, MO, USA

**Jessica A. R. Logan** The Crane Center for Early Childhood Research and Policy, Ohio State University, Columbus, OH, USA

**D. Betsy McCoach** University of Connecticut, Storrs, CT, USA

**Kristen L. McMaster** University of Minnesota, Minneapolis, MN, USA

**Alison M. Mitchell** Lexia Learning, Concord, MA, USA

**Nancy Nelson** Center on Teaching and Learning, University of Oregon, Eugene, TX, USA

**Peter M. Nelson** University of Minnesota, Minneapolis, MN, USA

**Ethan R. Van Norman** University of Minnesota, Minneapolis, MN, USA

**Miriam Ortiz** Southern Methodist University, Dallas, TX, USA

**David C. Parker** ServeMinnesota, Minneapolis, MN, USA

**Jill M. Pentimonti** The Crane Center for Early Childhood Research and Policy, Ohio State University, Columbus, OH, USA

**Yaacov Petscher** Florida Center for Reading Research, Florida State University, Tallahassee, FL, USA

**Apryl Poch** University of Missouri, Columbia, MO, USA

**John J. Prindle** Max Planck Institute, Berlin, Germany

**Cynthia S. Puranik** University of Pittsburgh, Pittsburgh, PA, USA

**Kristen D. Ritchey** University of Delaware, Newark, NJ, USA

**Kristi L. Santi** College of Education, University of Houston, Houston, TX, USA

**Lina Shanley** Center on Teaching and Learning, University of Oregon, Eugene, OR, USA

**Lina Shanley** Center on Teaching and Learning, University of Oregon, Eugene, TX, USA

**Benjamin Silbergitt** Technology and Information Educational Services, Falcon Heights, MN, USA

**Keith Smolkowski** Oregon Research Institute, Eugene, Oregon, USA

**Lisa Strycker** Oregon Research Institute, Eugene, Oregon, USA

**Huihui Yu** University of Connecticut, Storrs, CT, USA



# Chapter 1

## What Is Fluency?

Gina Biancarosa and Lina Shanley

Fluency is a deceptively simple term, but one that differs in its connotations depending on the literature or field referenced. Even within a single field in education, the term's meaning is sometimes debatable. Although the meaning of *fluency* invariably touches on its earliest usage in English in the early seventeenth century, when it meant an unrestrained or smooth flow (fluency, n.d.), the peculiar connotations within each field can vary from incredible specificity to equally incredible generality. This chapter presents definitions of and perspectives on fluency from three educational fields: language acquisition, reading, and mathematics. It then explores the foundations of the concept of fluency in research and finally makes recommendations for improved clarity regarding the term across fields.

### Fluency in Language Proficiency

In casual conversation, the term *fluency* tends to be used most frequently in reference to language proficiency, especially second language (L2) proficiency (e.g., “she is a fluent Spanish speaker”). Although in common terms L2 fluency conveys both ease and accuracy in speaking (fluency, n.d.), in the field of linguistics fluency has a more specific, distinct meaning. In recent definitions of linguistic fluency, accuracy or correctness in expression is distinctly *not* a part of fluency (Housen, Kuiken, & Vedder, 2012). Instead, fluency in language proficiency is “the ability to produce the L2 [second language] with native-like rapidity, pausing, hesitation, or reformulation” (p. 2). The distinction between fluency and accuracy, as well as complexity, has been bolstered by factor analyses that identify each as uncorrelated

---

G. Biancarosa (✉) · L. Shanley  
Center on Teaching and Learning, University of Oregon, Eugene OR, USA  
e-mail: ginab@uoregon.edu

L. Shanley  
e-mail: shanley2@uoregon.edu

© Springer Science+Business Media, LLC 2016  
K. D. Cummings, Y. Petscher (eds.), *The Fluency Construct*,  
DOI 10.1007/978-1-4939-2803-3\_1

factors in L2 performance (e.g., Norris & Ortega, 2009). Note that complexity in this field refers to have more elaborated knowledge of a language's syntax and lexicon (Housen et al., 2012).

Despite the distinction between fluency and accuracy or complexity, fluency can also be used within the linguistics field to refer more broadly to global L2 proficiency, a fact bemoaned by many (e.g., Derwing, Rossiter, Munro, & Thomson, 2004; Housen et al., 2012; Schmidt, 1992). At issue is that use of a more general meaning for the term (i.e., as "broadly synonymous with language mastery and native-like performance" (Chambers, 1997, p. 536)) and the predominance of its use in casual conversation generate confusion in research and in practice. For example, while psycholinguistic researchers emphasize grammatical knowledge in their definition of fluency, researchers in communications emphasize communicative competence which involves far more than just grammar (Schmidt, 1992). These multiple meanings of the term fluency become engrained in the minds of those who rely on both sources of research to inform their work, be that work research- or practice-oriented, leading to more and more reliance on the holistic meaning (Chambers, 1997; Schmidt, 1992). As a result, many have argued for restricting the use of the term fluency in speech to easily quantifiable aspects of speech related to timing (Chambers, 1997) and specifically to procedural language skills that can become automatic (Derwing et al., 2004; Schmidt, 1992). Thus, while fluency in language acquisition is distinct from accuracy and complexity, these aspects of linguistic proficiency seem to serve as prerequisites for fluency in that they must be automatic for fluent, native-like L2 speech to occur.

At the same time, fluency in L2 proficiency, along with accuracy and complexity, is not a fixed state. Rather, it is affected by the task at hand and the extent to which planning is afforded (e.g., Derwing et al., 2004; Foster & Skehan, 1996, 2013; Schmidt, 1992; Yuan & Ellis, 2003). For example, novice L2 speakers perform better on conversational tasks than on more formal speech tasks like telling a story (Derwing et al., 2004). In addition, given time to plan speech improves the fluency (and complexity) of L2 speakers (Yuan & Ellis, 2003). Moreover, task and planning time interact in their effects on fluency and other aspects of L2 proficiency (Foster & Skehan, 1996).

Contrast this nuanced definition and understanding of L2 linguistic fluency with the definition of native language (L1) fluency. According to Fillmore (1979), L1 fluency can refer to as many as four different definitions, all of which are relatively widely accepted and are used interchangeably as though they represent no important difference in meaning. One definition of L1 fluency is speed and ease in speech. Second is quality of speech, as in complexity and coherence of utterances. The third definition focuses on pragmatics, or the social niceties of language, such that L1 fluency represents an ability to speak appropriately in a variety of social contexts. The last definition relates to the ability to play with language, with fluency being the ability to manipulate language in creative ways, as in the creation and use of jokes, puns, metaphors, analogies, irony, etc. This multiplicity of definitions in many ways reflects earlier disagreements among theorists about how to define fluency in L2 acquisition. It may also reflect the relative scarcity of research on L1 fluency in and of itself outside of specialized fields

like speech and language pathology. In the field of speech and language pathology, the definition of fluency endorsed by the American Speech–Language–Hearing Association (ASHA) invokes not one, but at least two of the above-mentioned definitions. ASHA (1999) defines fluency as “the aspect of speech production that refers to the continuity, smoothness, rate, and/or effort with which phonologic, lexical, morphologic, and/or syntactic language units are spoken.” ASHA also notes the confusion surrounding the term fluency and that the roots of this confusion are multifaceted, arising from historical, cultural, linguistic, and practical sources.

## Fluency in Reading

Debates about the meaning of “fluency” are not unique to the language proficiency field. The field of reading has also long debated its definition and nature, and these debates have intensified recently. Interest in fluency in reading dates back to a seminal work by LaBerge and Samuels (1974) in which they theorized that processes in reading developed in two stages: the accuracy stage, in which attention is necessary to successful performance, and the *automaticity* stage, in which attention is no longer necessary to successful performance. Moreover, they detail a model wherein words presented in writing are processed in working memory first visually, then phonologically (although this phase can sometimes be skipped), then semantically, and finally stored in episodic memory, where a mental account of what is being read is developed. They invoke the concept of fluency in detailing their theory by noting that “the goal of fluent reading” is that “the reader can maintain his attention continuously on the meaning units of semantic memory, while the decoding from visual to semantic systems proceeds automatically” (p. 313). According to their theory, automatization of decoding is necessary for a reader to be considered fluent. Moreover, their theory explains why, as Smith and Holmes (1971) noted, a reader who is not fluently decoding is “not going to comprehend what he is reading simply because his memory system will not be able to retain, organize, and store the fragmentary information in any efficient way” (p. 412).

LaBerge and Samuels (1974) do not directly define oral reading fluency. Their theory is explicitly one of automatic reading processes. Nonetheless, they do note that to the proficient reader reading feels like an integrated, effortless process in which a variety of subskills (e.g., letter recognition, letter-sound associations, blending, etc.) are mastered to an automatic level such that the reader is not even aware of using and integrating subskills when reading. As long as words are recognized and their meanings are accessed automatically, a reader remains focused on deriving and retaining meaning at the episodic level. They further note that their model accounts for *word callers* in that there are readers who read aloud well without directing attention to the semantics of what they read, as well as for the phenomenon of a proficient reader who cannot recall what has been read because of an inattention to episodic processing of text. They explain word callers by suggesting that they are early readers who focus their attention solely on decoding (converting

visual representations into phonological ones), whereas the inattentive proficient reader decodes and activates word meanings so automatically that it literally leaves the mind “free to wander to other matters” (p. 320). In the latter case, the reader is not using attentional resources to organize automatically activated meanings into any higher-order structure.

Since the time of LaBerge and Samuels’ (1974) casual reference to word callers, their existence has been hotly debated. Some question whether word calling truly occurs (Hamilton & Shinn, 2003; Meisinger, Bradley, Schwanenflugel, & Kuhn, 2010; Meisinger, Bradley, Schwanenflugel, Kuhn, & Morris, 2009; Nathan & Stanovich, 1991). Others claim it is an unintended negative consequence of too much instructional and assessment attention to phonics and fluency (e.g., Goodman, 1973; Samuels, 2007). Whether or not one believes, as do Nathan and Stanovich (1991), that the phenomenon of word callers is a red herring, the forgetting after reading phenomenon cited by LaBerge and Samuels is uncontested (perhaps due to most proficient readers having experienced it firsthand). What is unclear from their references to these phenomena is whether LaBerge and Samuels would consider word calling and reading without recall instances of *fluent* reading, because their criteria for fluency are not explicitly defined.

Within less than a decade of the publication of LaBerge and Samuels’ theory, oral reading fluency became a frequently used indicator of automaticity in reading (e.g., Deno, Mirkin, & Chiang, 1982; Fuchs, Deno, & Marsten, 1983) and fluency interventions became a targeted subject of study (e.g., Allington, 1983; Martin & Meltzer, 1976; Samuels, 1979). Along the way, the role of prosody in the definition of reading fluency became contested (e.g., Allington, 1983), despite the fact that it played no explicit role in many of the original theories regarding reading fluency (e.g., LaBerge & Samuels, 1974; Nathan & Stanovich, 1991; Perfetti, 1985). Prosody refers to a reader’s ability to alter vocal volume and pitch and utilize pauses to scaffold and convey meaning when reading aloud (Benjamin & Schwanenflugel, 2010). The National Reading Panel seemingly settled the debate by defining oral reading fluency as reading “with speed, accuracy, and proper expression” (NICHD, 2000, pp. 31), and a special 2002 National Assessment of Educational Progress study further bolstered the argument that prosody was an essential component of reading fluency (Daane, Campbell, Grigg, Goodman, & Oranje, 2005). Recent research in oral reading fluency has specifically attempted to disentangle prosody from rate and accuracy (Benjamin, Schwanenflugel, Meisinger, Groff, Kuhn, & Steiner, 2013; Cowie, Douglas-Cowie, & Wichmann, 2002). These studies have found that readers can be fast and accurate but not necessarily prosodic in their reading; however, prosody rarely exists in the absence of accuracy and speed. Nonetheless, most commonly used measures of oral reading fluency, such as curriculum-based measures (CBMs), focus solely on accuracy and speed in reading and largely continue to ignore prosody (e.g., Deno et al., 1982; Fuchs et al., 1983; Fuchs, Fuchs, Hosp, & Jenkins, 2001).

Although the privileging of speed and accuracy in reading fluency measurement is at least partially due to the comparative ease with which speed and accuracy are measured in comparison to prosody (e.g., Fuchs et al., 2001), some have argued that

the neglect of prosody in fluency measures reflects a theoretical misunderstanding or misspecification of reading development (e.g., Allington, 1983; Slocum, Street, & Gilberts, 1995). Allington (1983) has argued that poor oral reading fluency is erroneously interpreted as an indicator (or symptom) of poor efficiency in word recognition, in other words, nonautomatic decoding. Allington joins Schreiber (1980) in arguing that merely learning to decode quickly and correctly does not fully explain fluent reading. Specifically, they cite early fluency intervention research (Samuels, 1979) in which the effects of repeated reading of connected texts are contrasted with the effects of training in automatic recognition of individual words. They argue that the superior effects of repeated reading suggest more is at issue than simply achieving decoding efficiency and that developing efficiency in the parsing of syntax also contributes to reading fluency.<sup>1</sup> This *syntactic processing* of text, reading not so much word-by-word as in meaningful word groups, supports a reader's ability to comprehend a text as a whole (i.e., text-level semantic processing). Recent research on the development of fluency supports this notion (e.g., Klauda & Guthrie, 2008; Kuhn & Stahl, 2003; Miller & Schwanenflugel, 2008; Rasinski, Rikli, & Johnston, 2009). Nevertheless, it remains unclear whether the original LaBerge and Samuels' theory of automaticity in reading accounts for this development. Syntactic processing might constitute another intermediate phase of reading not accounted for in the original model, or it could be argued as occurring simultaneously and integrally with semantic processing.

## Fluency in Mathematics

The definition of fluency in the field of mathematics also suffers from some ambiguity. There is general agreement that performance on complex mathematical tasks depends on the ability to quickly and accurately carry out the necessary operations (Bull & Johnston, 1997; Floyd, Evans, & McGrew, 2003). Indeed, Floyd et al. (2003) defined processing speed as the "ability to perform simple cognitive tasks quickly, especially when under pressure to maintain focused attention and concentration" (p. 159). Yet, much like the role of syntactic processing in oral reading fluency is not fully specified, what precisely must become both quick and accurate is not fully determined. For example, researchers have investigated the role of general processing speed in mathematics proficiency, finding correlations with mathematics difficulties that suggest a deficit in general processing speed at least partially explains mathematics difficulties (Bull & Johnston, 1997). But more recent research investigating the degree to which processing speed affects mathematics performance has found conflicting evidence for a direct effect (Floyd et al., 2003) versus an indirect, mediating effect (Rindermann & Neubauer, 2004). Similarly, evidence

---

<sup>1</sup> Interestingly, Samuels has since become one of the more vocal proponents of the importance of not only prosody, but also comprehension in the definition and measurement of oral reading fluency (e.g., Samuels, Ediger, & Fautsch-Patridge, 2005; Samuels, 2007).

on the role of item identification (e.g., tracking the operational signs in a fact fluency task)—which is parallel to the concept of phonological recognition of words in reading—has been mixed. Although some research points to a crucial role for quick and accurate item identification in successful computational performance because items that are identified more quickly free mental resources for computation (e.g., Case, Kurland, & Goldberg, 1982), other research suggests that experimentally disrupting item identification (i.e., enforcing slower identification of items) does not disrupt successful computational performance (e.g., Deschuyteneer & Vandierendonck, 2005).

Due to the more obvious generative nature of mathematics problem-solving (i.e., the student must generate a visible or audible answer to a problem, which is more akin to writing than to reading), theories articulate another important aspect of mathematics fluency: response selection, which is defined as “a process that is required in order to assimilate the different response alternatives and to select a correct response between activated alternatives” (Deschuyteneer & Vandierendonck, 2005, p. 1473). For example, one must be able to not only identify problem type but also attend to details such as operational sign (e.g., +, ÷) in order to determine an appropriate range of responses from which to select one’s own response. The closest parallel to this part of mathematics fluency in reading fluency might be when the reader must choose among meanings for a polysemous word (i.e., a word with many meanings; e.g., *bank*) or even decide on whether a vowel within a word uses a short, long, or other sound. A more intuitive parallel can be construed in linguistic fluency, perhaps because of its emphasis on speech production. Specifically, response selection in mathematics fluency can be seen as paralleling accuracy in L2 proficiency because of the requirement of considering and choosing the best among many alternatives. Thus, fluency must be tempered based on the task demand, or else accuracy may be sacrificed for speed.

Similar to reading fluency, fluency in mathematics has become a strong focus in assessment and instruction (e.g., Carnine, 1997; Clarke, Nelson, & Shanley, *in press*; Gersten, Jordan, & Flojo, 2005; Rhymer Dittmer, Skinner, & Jackson, 2000). Researchers have noticed that students who struggle with mathematics not only have difficulty with both accuracy and speed in responding to mathematical problems (Gersten et al., 2005; Hasselbring, Goin, & Bransford, 1988), but also struggle with efficient and effective strategy application (e.g., Geary, 1993; Geary, Brown, & Samaranayake, 1991; Gersten et al., 2005). As a result, timed measures of mathematical skills have been employed in the early identification of students at risk for poor progress in mathematics (e.g., Chard, Clarke, Baker, Otterstedt, Braun, & Katz, 2005; Clarke & Shinn, 2004). Nonetheless, some argue that this attention to fluency in mathematics fails to differentiate the real root of students’ struggles with mathematics; whether that is a general deficit in processing speed or a specific deficit in mathematical understanding (Chiappe, 2005). That is, the lack of mathematical fluency may represent not so much lack of speed and accuracy in mathematics procedures as a fundamental deficit in the understanding of numbers.



## Fluency as Automaticity in Procedural Skill

As the preceding sections have hopefully made clear, debate around what constitutes fluency is not unique to any particular field in education. How then are we to arrive at a precise definition of fluency? A return to the origins of fluency as a concept in education may help.

The most consistent concept featured in each of the reviews above is that fluency is connected to automaticity. Again and again the idea that skills (and usually procedural skills) must become not only accurate but also automatic is invoked (Bull & Johnston, 1997; Derwing et al., 2004; Floyd et al., 2003; LaBerge & Samuels, 1974; Schmidt, 1992). Although he was defining fluency in language, Schmidt highlights this critical component by calling fluency “automatic procedural skill” (p. 358). But really, what does it mean to be automatic? And why procedural skill?

### *Automaticity*

Bargh and Chartrand (1999) argue that what all definitions of automaticity share is a clear definition of what automaticity is *not*, rather than of what it is. The main thing that automatic processes are not is *conscious*. Conscious processes are ones of which an individual agent is *aware*, that the agent *intends*, that require *effort* by the agent, and that are actively *controlled* by the agent. In contrast, automatic processes are ones in which the processes require no or limited attention, intention, effort, and control. As a result, conscious mental capacity is freed to attend to other things.

To understand just how unconscious automatic processes can be a common example given is the commute from home to work and back again. For those with routinized and dependable schedules, once the commuting procedure is mastered, the commuting process can become so automated that we have no conscious memory of anything specific to that process that occurred along the way, unless it deviates from the usual. An even simpler example is how we respond to reading the phrase “don’t think of a pink elephant.” Did you think of a pink elephant? Chances are, you did. At issue here is that even though the phrase directs us *not* to think of something, proficient readers cannot follow this directive because the words “pink” and “elephant” are not challenging for us to decode and their meanings are immediately activated whether we want them to be or not. Thus, we think of a pink elephant. Inevitably, reading words conjures up their meanings and associated mental images so long as decoding the words is not challenging and the words’ meanings are sufficiently well known.

Another important aspect of automaticity is that this unconscious execution of processes can be acquired either consciously or unconsciously (Bargh & Chartrand, 1999). In the commuting example above, the acquisition of automaticity is likely unconscious. Once one has a route and routine, practice does the trick. As Bargh and Chartrand (1999) explain, unconscious acquisition of automaticity has only two “necessary and sufficient ingredients [...] frequency and consistency of use of the same set of component mental processes under the same circumstances” (p. 469). For con-

scious acquisition of automaticity, one merely needs to set out intentionally to do the same. For example, if you wish to become more automatic at riding a bike, you can decide to practice regularly. Assuming no impediments to your ability to ride a bike, repeated intentional practice should yield more automatized bike riding. One sets out intentionally to practice a process until successful execution no longer requires conscious, intentional control and attention becomes superfluous to its execution.<sup>2</sup>

Yet another important part of automaticity is that when a complex process, such as commuting, is involved, it is better applied to components of the process rather than holistically. Although the commuting process, such as driving to and from work, may seem simple enough at first blush, it actually involves the coordination of a number of component processes. Driving to and from work requires not only a memorization of the route, but a coordination of other activities so as to avoid being late and of course also knowledge of how to drive a car. When rush hour is involved, coping with increased traffic adds to the necessary skill set. And when carpooling is involved, the driver also needs to be able to cope with conversation while driving. Any one of these components of the commuting process can be automatized ... or not. While speed does not figure explicitly into a definition of automaticity, a clear consequence of lack of automaticity is temporal: the agent engaged in the process hesitates, slows down, and stutters in responding to tasks for which automaticity has not been achieved. Hence, the unpracticed carpooling commuter may halt speaking mid-conversation each time turning is required or need to turn down the radio in order to parallel park. In contrast, the practiced and proficient commuter will tend to show no sign of disruption or competition for cognitive resources despite changes in route, conversational topic, traffic, or other attentional demands. Given the obvious coordination of many component processes in complex skills, some have argued that automaticity of cognitive processes of any sort is best understood by first completing a comprehensive and precise specification of all the component processes involved (Jonides, Naveh-Benjamin, & Palmer, 1985).

### ***Procedural Skill***

According to John R. Anderson (1982; 1996), knowledge has two basic forms: declarative and procedural. Declarative knowledge is *knowing that*, whereas procedural knowledge is *knowing how*. Automaticity is said to develop only for this latter type of knowledge (Logan, 1985). Even within the concept of procedures becoming automatic, theorists emphasize components of tasks, rather than tasks as a whole, as what become automatized (e.g., LaBerge & Samuels, 1974; Logan 1985).

Take our earlier commuting example. According to cognitive psychology, we would be better off describing each component of the commuting process and what automaticity in it looks like rather than describing the process as a whole. These com-

---

<sup>2</sup> Hasher and Zacks (1979) argue that automaticity should *not* be applied to processes that benefit from practice, but this perspective is decidedly in the minority (for discussion see Fisk & Schneider, 1984).



ponents do not refer to steps in the commute, but rather to each of the processes upon which the execution of the daily commute relies. Some of the processes involved include but are not limited to conversing, retrieving route knowledge, and driving, each of which naturally could be broken into several component processes of its own (Hoover & Gough, 1990). Similarly, in the field of reading, for example, the reading process can be broken into at least two processes (i.e., decoding and comprehension), each of which also has subprocesses that are required for skillful execution. For example, skilled comprehension relies on automaticity in many mental processes (Thurlow & van den Broek, 1997). Among these is the use of comprehension strategies such as summarizing, predicting, and monitoring one's comprehension. The skilled reader will demonstrate automaticity in each of the component processes of reading.

Nonetheless, automaticity is not purely synonymous with skill in the execution of procedural knowledge (Logan, 1985). Logan argues that skills, like speaking or reading, "consist of collections of automatic processes that are recruited to perform the skilled task" (p. 368) but that the difference between skilled and unskilled performances cannot be ascribed to automaticity alone. Instead, Logan suggests that skilled performers are also likely to have more declarative knowledge than unskilled performers, as well as more understanding (or, more specifically, metacognition) that allows them to recruit that *declarative* knowledge for more efficient and effective use of *procedural* knowledge. In short, Logan argues that skill is defined by the coordinated execution of automatized procedural knowledge in the presence of specific goals and constraints. As a less complex example, consider running. We learn to run relatively early in life and automatize running fairly rapidly. Nonetheless, not every runner is a skilled runner. Differences in skill not only relate partially to practice but also depend in large part on the task set for the runner. For example, it is the rare runner who is equally skilled at sprint running, marathon running, relay running, and obstacle running (e.g., hurdles). Each type of running requires specific foci for practice. While the same essential component processes are involved in the act of running, each differs in the other processes with which running must be coordinated (e.g., hand offs in relay running, jumping in obstacle running). Whereas some processes inevitably overlap, others do not and may even conflict. Thus, skill in running overall relies on the runner's ability to adapt the execution of automatized processes contingent on the demands of the task at hand, to coordinate both declarative and procedural knowledge for an optimal outcome.

As a result, Logan (1985) also suggests that assessing automaticity for a specific process (e.g., running pace, stride length) in relative isolation from other processes or specific tasks is not enough and may leave us with an incomplete picture of skill. When processes are assessed out of context, an important aspect of skill in performance is lost, namely the more skilled performer's ability to anticipate responses based on context. Take, for example, research that has demonstrated that speed and accuracy in reading words depends on whether the words are presented in context (e.g., a passage or story) or out of context (e.g., in a list or a passage or story with the words in a random order; Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003). In the study by Jenkins and colleagues (2003), fourth graders read more quickly, by an average of 44 correct words or more per minute, when reading words in context than out of context. Moreover, speed of reading in context better predicted reading

comprehension on a separate criterion measure of reading comprehension than it did either out of context speed measure, suggesting that the more closely aligned a component process is to the fully executed target skill (here, reading comprehension), the more informative of the skill it is.

Thus, if we are to truly understand automaticity, not only of component skills but also of their coordinated use, Logan argues that we must study it in both situations, in and out of context. To continue the reading example, assessing decoding both in and out of context has the potential to yield information. Reading words out of context isolates the component skills of decoding allowing insight into the automaticity of decoding alone,<sup>3</sup> while reading words in context comes closer to the target skill allowing insight into the coordination of decoding with other processes, including syntactic and semantic processing.

## *Fluency*

If we define fluency as involving automaticity in procedural skill, then it becomes necessary to define more precisely the procedural skills and the signs of automaticity in them. This issue of definition is exemplified in an article by Samuels (2007), one of the originators of automaticity theory relative to reading. In this article, he contends that reading fluency measures such as the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) are not truly fluency measures, rather he argues:

[N]one of the DIBELS instruments are tests of fluency, only speed .... To understand the essential characteristic of fluency and what its window dressings are, we must look to automaticity theory for guidance (LaBerge & Samuels, 1974). At the risk of oversimplification, in order to comprehend a text, one must identify the words on the page *and* one must construct their meaning. If all of a reader's cognitive resources are focused on and consumed by word recognition, as happens with beginning reading, then comprehension cannot occur at the same time. (Samuels, 2007, p. 564)

Based on the preceding sections regarding automaticity theory, Logan (1985) and others might argue that Samuels and the authors of DIBELS are defining automaticity for different procedural skills or under different constraints. DIBELS assessments focus on the component process of decoding (and other component skills) under different conditions (e.g., word reading out of context, word reading in context), while Samuels is insisting on assessments that examine decoding with tasks that require readers to read for meaning and not just demonstrate skill in isolated procedures. For Samuels, fluency in reference to reading necessarily invokes coordinated performance of all the component processes involved in reading. But for assessments like DIBELS, fluency can refer to any of the myriad component processes involved in reading.

---

<sup>3</sup> Note that among reading researchers interested in full isolation of decoding, reading of decodable nonsense words out of context is considered preferable to reading real words out of context because it isolates decoding even from sight recognition and effects of vocabulary knowledge (e.g., Ehri, 2005).

## Recommendations for the Study of “Fluency”

Based on the preceding, we posit that it is time to move beyond fluency as a term. Too often fluency draws up notions of complex, holistic skills. The confusion is only natural, given the origins of the term in the Latin verb *fluere* meaning *to flow* (fluency, n.d.). Nonetheless, as long as any number of researchers and theorists treats fluency as a unitary concept within a field, much time and effort will be lost arguing about whether and what figures into this vague concept which we call fluency. Instead of fluency, we recommend in this section that using the term automaticity invokes far less confusion.

Disagreement about the meaning of fluency can lead to debates about its role in complex processes that are pointless because the underlying meaning of the term is arguably what is at issue. For example, in the field of reading, the idea that reading can be broken into the two primary component processes of decoding and comprehension is called the “simple view of reading” (Gough & Tunmer, 1986; Hoover & Gough, 1990). However, there is a long-standing debate that has only picked up in intensity recently about whether fluency ought to be included as a third primary process in this and other theoretical models. Research has not yielded a clear answer (e.g., Adlof, Catts, & Little, 2006; Kershaw & Schatschneider, 2012; Silverman, Speece, Harring, & Richey, 2013; Tilstra, McMaster, van den Broek, Kendeou, & Rapp, 2009). In some cases, fluency stands out a separable construct, but in others it does not. Discrepancies in findings might be attributed to differences in analytic approach, but perhaps a more important difference across such studies is in the measures used. For instance, in two recent studies measures of word list reading fluency were included as indicators of fluent reading (Adlof et al., 2006; Silverman et al., 2013), but in two others they were not (Kershaw & Schatschneider, 2012; Tilstra et al., 2009). Moreover, one study included measures of word list reading fluency as measures *not* of fluency but of decoding (Kershaw & Schatschneider, 2012). Notably, none of these studies included fluency measures tapping prosody as an aspect of fluency. We would argue that the inconsistency in how fluency is represented and the inclusion of fluency measures as measures of decoding, which are not unique to this group of studies, reflect a fundamental lack of clarity and agreement in the field as to what fluency is.

What if instead of debating what is or is not fluency, we followed the rather old advice of Jonides et al. (1985) to focus on automaticity and identify first the components of more complex processes and then set the criteria for automaticity rooted in those component processes rather than in more complex behavior? If we did, within each field, be it language, reading, or mathematics, we would attempt to lay out each of the component processes involved in skilled performance and then determine what criteria we will use for detecting automaticity in those processes rather than more holistically in language, reading, and mathematics. In such an approach, for our reading theorists debating the role of “fluency,” we would instead investigate the role of automaticity in reading words, which would include the reading of words both in and out of context, and automaticity in other reading processes as well. For the growing number of researchers who argue a role for prosody in proficient reading (Allington, 1983; Benjamin & Schwanenflugel,

2010; Klauda & Guthrie, 2008; Kuhn & Stahl, 2003; Miller & Schwanenflugel, 2008; Rasinski et al., 2009; Samuels, 2007), we would also investigate automaticity in prosody in reading aloud. In other words, we propose that decoding can be construed as one process and prosodic parsing of text as another and that fluency be used in reference to neither. Instead, we suggest the field define how automatization of each process can be construed separate from the other. We say this despite the fact that each inevitably interacts with the other in the execution of reading, and all the more so the longer and more authentic the text being read.

Moreover, our position is not a new one and applies to each of the fields covered in our review. Although usually cited as support for the role of prosody in fluency, Hudson, Pullen, Lane, and Torgesen (2009) suggested that not only can reading fluency be broken down into at least four components (one of which is prosody) but also that one of those components (i.e., decoding fluency) can be broken down into further subcomponents of fluency. In many ways, the field of L2 acquisition has made the most progress in this vein, at least in terms of the acceptance of the idea that separating accuracy and complexity of utterances are and should be separable from fluency in utterances. The acronym *CAF*, which stands for complexity, accuracy, and fluency, has been used quite widely in the literature on L2 acquisition since the 1990s (e.g., Housen & Kuiken, 2009). Yet, even within this field there have been calls for greater specificity for each of the constructs making up CAF and calls for addressing the multidimensional nature of fluency (e.g., Pallotti, 2009). In other words, even within L2 acquisition theory and research, the use of that pesky term *fluency* still invites confusion. If fluency in language proficiency is “the ability to produce the L2 [second language] with native-like rapidity, pausing, hesitation, or reformulation” (Housen et al., 2012, p. 2), then there remains a need to define and investigate each of these components of fluency in language proficiency (Pallotti, 2009). Critical to clarity here, as in the field of reading, is disentangling speed (i.e., rapidity) from prosody (i.e., pausing) and from metacognitive monitoring and repair practices.

At the very least, if we follow the lead of L2 acquisition, accuracy is its own separable construct and should not figure into measures of automaticity. Thus, for example, in the reading field the common approach of using rate of accurate reading (i.e., correct words per minute) as an index of automaticity confounds two aspects of reading proficiency: accuracy and speed. Although this practice offers some efficiencies in measurement and statistical modeling and is the most common form of measuring oral reading fluency (Deno et al., 1982; Fuchs et al., 1983; Fuchs et al., 2001; Jenkins et al., 2003; see also Espin & Deno, *in press*), the practice does not always translate well to all contexts. For example, although Jenkins et al. (2003) found that accuracy (measured as proportion of correct words to total words read) did not add anything over and above speed (measured as correct words per minute) to the explanation of reading comprehension, the two measures were not truly independent of each other. Even when accuracy is measured as number of errors per minute (e.g., Baker et al., *in press*), accuracy, or here the lack of accuracy, is still confounded with rate.

Despite the seemingly inevitable confound between accuracy and rate when they are derived from the same prompts, recent research has demonstrated that even accuracy can uniquely contribute to the explanation of reading comprehension, particularly in contexts where speed of reading is not as widespread and variability in accuracy is greater. For example, studies have found that measuring accuracy as well as speed yields more information about readers among students with disabilities (e.g., Puolakanaho et al., 2008) or in contexts where universal literacy is further from realization (e.g., Farukh & Vulchanova, 2014; Vagh & Biancarosa, 2011). Nonetheless, the reader who reads *quickly but inaccurately* is a rarity. If readers like that exist at all, they have yet to turn up in a wide range of studies examining heterogeneity among the struggling readers at varying grade levels (e.g., Brassuer-Hock, Hock, Kieffer, Biancarosa, & Deshler, 2011; Buly & Valencia, 2002; Catts, Hogan, & Adlof, 2005).

For the vast majority of performers (in reading or otherwise) there is a relation between accuracy and speed (e.g., MacKay, 1982). As accuracy improves, speed also tends to improve. In other words, until accuracy reaches some level of proficiency, speed cannot improve very much. However, this phenomenon relies on practice as a mechanism for improvement of both accuracy and rate, and more importantly that speed is not intentionally valued over accuracy. In fact, in the latter case efforts to execute a process at either much greater or much slower rates than are typically practiced will result in compromised accuracy. Although the evidence regarding this phenomenon have been demonstrated primarily in speech production and for physical processes (e.g., moving one's arm in a particular manner; see MacKay, 1982 for discussion), the link between accuracy and speed and their dependence on practice conditions are consistent with more general theories of automaticity (Bargh & Chartrand, 1999; Logan, 1985).

Despite the clear distinction between accuracy and speed, there may be practical reasons for confounding the two in measurement. In fields like education, the idea of measuring for speed without also measuring for accuracy risks a lack of face validity, wherein a measure of just speed does not have the appearance of measuring the intended (or valued) construct. Although face validity plays no quantifiable role in measurement, it can have a profound impact on which measures get adopted as well as on test-taker behavior (Thorndike & Thorndike-Christ, 2010). Thus, particularly in educational practice if not in basic research, automaticity measures that confound accuracy and speed may be unavoidable.

Nevertheless, we argue that use of a term less fraught with confusion and debate than fluency is advisable. Instead of fluency, we suggest using the term automaticity itself or, alternatively, efficiency. Efficiency has the advantage of connoting not only speed but also accuracy. In other words, it invokes the idea of the speed-accuracy trade-off (MacKay, 1982) wherein speed without accuracy and accuracy without speed are of limited value. Perfetti was perhaps one of the first reading theorists to use the term efficiency when discussing theories of the reading process (e.g., Perfetti, 1985). More recently, measures of what has typically been called decoding fluency have also shunned the term fluency for efficiency

(i.e., Torgesen, Wagner, & Rashotte, 2012). Similar refinements of terminology exist within the field of mathematics as well, wherein accuracy and speed are defined separately and efficiency is seen as combining the two qualities in computational skill (e.g., Hoffman & Spataru, 2008). Indeed the president of the National Council of Teachers of Mathematics noted that “Focusing on efficiency rather than speed means valuing students’ ability to use strategic thinking to carry out a computation without being hindered by many unnecessary or confusing steps in the solution process” (Gojak, 2012). Thus, a shift in terminology is not necessarily an onerous task and could help to reduce confusion in practice and research in numerous fields.

## Conclusions

So what is fluency? It depends on whom you ask. Even within fairly narrow fields, definitions vary. Thus, we recommend treating the term fluency as a holistic description of a skilled performance. In addition, we recommend reserving more specific terms for processes that are components of skilled performance. These recommendations are by no means revolutionary, but as our review of the literature in three education fields demonstrates they have yet to catch on in studies of fluency in language acquisition, reading, and mathematics. The advantage of using more precise terms for the component processes involved in complex procedural skills like speaking in a foreign language, reading, or solving a mathematics problem is an avoidance of age-old confusion and debates that have at their root the term “fluency.” We should cease arguing about fluency and what it means and instead focus on each component process involved in complex procedural skills and the characteristics of proficiency in each process (e.g., accuracy, speed, efficiency, and complexity). This improved clarity would also serve to support efforts to understand the complex interactions between task and proficiency (i.e., how and why individuals might display superior accuracy or speed in one context than in another). An emphasis on clarity and specificity in terminology can also serve to inform efforts to improve fluency as a term by forcing developers and evaluators of instruction and intervention to reconsider what precisely they intend to affect (i.e., overall proficiency in a complex skilled performance or automaticity in some component aspect of a skilled performance). Most importantly, without a renewed commitment to precision in the terms we use, we risk wasting more time splitting hairs about what fluency is and who should define it. We would argue that whatever the field, the public is better served by and our attention is better turned toward developing nuanced and testable theories, measures, and interventions regarding the skills in which we are interested.



## References

- Adlof, S. M., Catts, H. W., & Little, T. D. (2006). Should the simple view of reading include a fluency component? *Reading and Writing, 19*(9), 933–958.
- Allington, R. L. (1983). Fluency: The neglected reading goal. *Reading Teacher, 36*(6), 556–561.
- American Speech-Language-Hearing Association. (1999). Terminology pertaining to fluency and fluency disorders: guidelines (Guidelines). [www.asha.org/policy](http://www.asha.org/policy). <http://www.asha.org/docs/html/GL1999-00063.html#sthash.7J4aMThZ.dpuf>.
- Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369–406.
- Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist, 51*(4), 355–365.
- Baker, D. L., Biancarosa, G., Park, B. J., Bousselot, T., Smith, J., Baker, S. K., Kame'enui, E. J., Alonzo, J., & Tindal, G. (in press). Validity of CBM measures of oral reading fluency and reading comprehension on high-stakes reading assessments in Grades 7 and 8. *Reading & Writing*. doi:10.1007/s11145-014-9505-4.
- Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54*(7), 462–479.
- Benjamin, R. G., & Schwanenflugel, P. J. (2010). Text complexity and oral reading prosody in young readers. *Reading Research Quarterly, 45*(4), 388–404.
- Benjamin, R. G., Schwanenflugel, P. J., Meisinger, E. B., Groff, C., Kuhn, M. R., & Steiner, L. (2013). A spectrographically grounded scale for evaluating reading expressiveness. *Reading Research Quarterly, 48*, 105–133. doi:10.1002/rrq.43.
- Brasseur-Hock, I. F., Hock, M. F., Kieffer, M. J., Biancarosa, G., & Deshler, D. D. (2011). Adolescent struggling readers in urban schools: Results of a latent class analysis. *Learning and Individual Differences, 21*(4), 438–452.
- Bull, R., & Johnston, R. S. (1997). Children's arithmetical difficulties: Contributions from processing speed, item identification, and short-term memory. *Journal of Experimental Child Psychology, 65*, 1–24.
- Buly, M. R., & Valencia, S. W. (2002). Below the bar: Profiles of students who fail state reading assessments. *Educational Evaluation and Policy Analysis, 24*(3), 219–239.
- Carnine, D. (1997). Instructional design in mathematics for students with learning disabilities. *Journal of Learning Disabilities, 30*(2), 130–141.
- Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology, 33*, 386–404. doi:10.1016/0022-0965(82)90054-6.
- Catts, H. W., Hogan, T. P., & Adlof, S. M. (2005). Developmental changes in reading and reading disabilities. In H. W. Catts & A. G. Kamhi (Eds.), *The connections between language and reading disabilities* (pp. 25–40). Mahwah: Lawrence Erlbaum.
- Chambers, F. (1997). What do we mean by fluency? *System, 25*(4), 535–544.
- Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention, 30*(2), 3–14.
- Chiappe, P. (2005). How reading research can inform mathematics difficulties: The search for the core deficit. *Journal of Learning Disabilities, 38*(4), 313–317.
- Clarke, B., & Shinn, M. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33*, 234–248.
- Clarke, B., Nelson, N., & Shanley, L. (in press). Mathematics fluency: more than the weekly timed test. In K. D. Cummings & Y. Petscher (Eds.), *Fluency metrics in education: Implications for test developers, researchers, and practitioners*. New York: Springer.
- Cowie, R., Douglas-Cowie, E., & Wichmann, A. (2002). Prosodic characteristics of skilled reading: Fluency and expressiveness in 8-10-year-old readers. *Language and Speech, 45*(1), 47–82.
- Daane, M. C., Campbell, J. R., Grigg, W. S., Goodman, M. J., & Oranje, A. (2005). *Fourth-grade students reading aloud: NAEP 2002 special study of oral reading (NCES 2006–469)*. U.S.

- Department of Education, Institute of Education Sciences, National Center for Education Statistics. Washington, DC: Government Printing Office.
- Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, 49, 36–45.
- Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54(4), 655–679.
- Deschuyteneer, M., & Vandierendonck, A. (2005). The role of response selection and input monitoring in solving simple arithmetical products. *Memory & Cognition*, 33, 1472–1483.
- Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, 9(2), 167–188.
- Espin, C. A., & Deno, S. L. (in-press). Oral reading fluency or reading aloud from text: An analysis through a unified view of construct validity. In K. D. Cummings & Y. Petscher (Eds.). *Fluency metrics in education: implications for test developers, researchers, and practitioners*. New York: Springer.
- Farukh, A., & Vulchanova, M. (2014). Predictors of reading in Urdu: Does deep orthography have an impact? *Dyslexia*, 20(2), 146–166.
- Fillmore, C. J. (1979). On fluency. In C. Fillmore, D. Kempler, & W. S.-Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85–101). New York: Academic.
- Fisk, A. D., & Schneider, W. (1984). Memory as a function of attention, level of processing, and automatization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(2), 181–197.
- Floyd, R., Evans, J. J., & McGrew, K. S. (2003). Relations between measures of Cathill-Horn-Carroll (CHC) cognitive abilities and mathematics achievement across the school-age years. *Psychology in the Schools*, 40, 155–171. doi:10.1002/pits.10083
- Fluency. (n.d.). In Oxford Dictionaries English online dictionary. <http://www.oxforddictionaries.com/definition/english/fluency>.
- Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18(3), 299–323.
- Foster, P., & Skehan, P. (2013). Anticipating a post-task activity: The effects on accuracy, complexity, and fluency of second language performance. *Canadian Modern Language Review*, 69(3), 249–273.
- Fuchs, L. S., Deno, S. L., & Marsten, D. (1983). Improving the reliability of curriculum-based measures of academic skills for psychoeducational decision making. *Diagnostique*, 8, 135–149.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5, 239–256.
- Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin*, 114(2), 345–362.
- Geary, D. C., Brown, S. C., & Samaranayake, V. A. (1991). Cognitive addition: A short longitudinal study of strategy choice and speed-of-processing differences in normal and mathematically disabled children. *Developmental Psychology*, 27(5), 787–797.
- Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, 38(4), 293–304.
- Gojak, L. M. (2012, Nov 1). Fluency: Simply fast and accurate? I think not! National Council of Teachers of Mathematics. <http://www.nctm.org/about/content.aspx?id=34791>
- Goodman, K. S. (1973). The 13th easy way to make learning to read difficult: A reaction to Gleitman and Rozin. *Reading Research Quarterly*, 8(4), 484–493.
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7, 6–10.
- Hamilton, C., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review*, 32(2), 228–240.
- Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, 108(3), 356–388.



- Hasselbring, T. S., Goin, L. I., & Bransford, J. D. (1988). Developing math automaticity in learning handicapped children: The role of computerized drill and practice. *Focus on Exceptional Children*, 20(6), 1–7.
- Hoffman, B., & Spatriu, A. (2008). The influence of self-efficacy and metacognitive prompting on math problem-solving efficiency. *Contemporary Educational Psychology*, 33(4), 875–893.
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, 2(2), 127–160.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency (CAF) in second language acquisition research (special issue). *Applied Linguistics*, 30(4), 461–473.
- Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy, and fluency: Definitions, measurement, and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp. 1–20). Philadelphia: John Benjamins.
- Hudson, R. F., Pullen, P. C., Lane, H. B., & Torgesen, J. K. (2009). The complex nature of reading fluency: A multidimensional view. *Reading & Writing Quarterly*, 25(1), 4–32.
- Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C.A., & Deno, S.L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95(4), 719–729.
- Jonides, J., Naveh-Benjamin, M., & Palmer, J. (1985). Assessing automaticity. *Acta Psychologica*, 60(2–3), 157–171.
- Kershaw, S., & Schatschneider, C. (2012). A latent variable approach to the simple view of reading. *Reading and Writing*, 25(2), 433–464.
- Klauda, S. L., & Guthrie, J. T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology*, 100(2), 310–321.
- Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, 95(1), 3–21.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293–323.
- Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology*, 39(2), 367–386.
- MacKay, D. G. (1982). The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review*, 89(5), 483–506.
- Martin, J. G., & Meltzer, R. H. (1976). Visual rhythms: Report on a method for facilitating the teaching of reading. *Journal of Literacy Research*, 8(2), 53–160.
- Meisinger, E. B., Bradley, B. A., Schwanenflugel, P. J., Kuhn, M. R., & Morris, R. D. (2009). Myth and reality of the word caller: The relation between teacher nominations and prevalence among elementary school children. *School Psychology Quarterly*, 24(3), 147–159.
- Meisinger, E. B., Bradley, B. A., Schwanenflugel, P. J., & Kuhn, M. R. (2010). Teachers' perceptions of word callers and related literacy concepts. *School Psychology Review*, 39(1), 54–68.
- Miller, J., & Schwanenflugel, P. J. (2008). A longitudinal study of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading Research Quarterly*, 43(4), 336–354.
- Nathan, R. G., & Stanovich, K. E. (1991). The causes and consequences of differences in reading fluency. *Theory into Practice*, 30(3), 176–184.
- National Institute of Child Health and Human Development. (2000). Report of the National Reading Panel. *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups (NIH Publication No. 00-4754)*. Washington, DC: U.S. Government Printing Office.
- Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578.
- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601.
- Perfetti, C. A. (1985). *Reading ability*. New York: Oxford.

- Puolakanaho, A., Ahonen, T., Aro, M., Eklund, K., Leppanen, P. H. T., Poikkeus, A.-M., Tolvanen, A., Torppa, M., & Lyytinen, H. (2008). Developmental links to very early phonological and language skills to second grade reading outcomes: Strong to accuracy but only minor to fluency. *Journal of Learning Disabilities, 41*(4), 353–370.
- Rasinski, T., Rikli, A., & Johnston, S. (2009). Reading fluency: More than automaticity? More than a concern for the primary grades? *Literacy Research & Instruction, 48*(4), 350–361.
- Rhymer, K. N., Dittmer, K. I., Skinner, C. H., & Jackson, B. (2000). Effectiveness of a multi-component treatment for improving mathematics fluency. *School Psychology Quarterly, 15*(1), 40–51.
- Rindermann, H., & Neubauer, A. C. (2004). Processing speed, intelligence, creativity, and school performance: Testing of causal hypotheses using structural equation models. *Intelligence, 32*, 573–589. doi:10.1016/j.intell.2004.06.005
- Samuels, S. J. (1979). The method of repeated readings. *Reading Teacher, 32*, 403–408.
- Samuels, S. J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency?: Commentary on “The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students.” *Reading Research Quarterly, 42*(4), 563–566.
- Samuels, S. J., Ediger, K.-A., & Fautsch-Patridge, T. (2005). The importance of fluent reading. *New England Reading Association Journal, 41*(1), 1–8.
- Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition, 14*(4), 357–385.
- Schreiber, P. A. (1980). On the acquisition of reading fluency. *Journal of Reading Behavior, 12*(3), 177–186.
- Silverman, R. D., Speece, D. L., Haring, J. R., & Ritchey, K. D. (2013). Fluency has a role in the simple view of reading. *Scientific Studies of Reading, 17*(2), 108–133.
- Slocum, T. A., Street, E. M., & Gilberts, G. (1995). A review of research and theory on the relation between oral reading rate and reading comprehension. *Journal of Behavioral Education, 5*, 377–398. doi: 10.1007/bf02114539.
- Smith, F., & Holmes, D. L. (1971). The independence of letter, word, and meaning identification. *Reading Research Quarterly, 6*, 394–415.
- Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed). Boston: Pearson.
- Thurlow, R., & van den Broek, P. (1997). Automaticity and inference generation during reading comprehension. *Reading and Writing Quarterly, 13*(2), 165–181.
- Tilstra, J., McMaster, K., van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. *Journal of Research in Reading, 32*(4), 383–401.
- Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). *Test of word reading efficiency* (2nd ed). Austin: Pro-Ed.
- Vagh, S. B., & Biancarosa, G. (2011, July). *Early literacy in Hindi: The role of oral reading fluency*. Poster presented at the 12th International Congress for the Study of Child Language, Montreal, Canada.
- Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics, 24*(1), 1–27.

## Part I

# Applied Use of Fluency Measures

In this first section, we focus on educational professionals who use fluency-based measurement data to make important educational decisions about students. The chapters in this section represent information regarding fluency metrics as applied to writing (Ritchey et al., Chap. 2, this volume) and mathematics measures (Clarke et al., Chap. 3, this volume) and also include syntheses of the work to date on using fluency-based measures to make universal screening (Lembke et al., Chap. 4, this volume) and response to intervention (Burns et al., Chap. 5, this volume) decisions.

Writing and mathematics metrics, as lesser studied curriculum-based measurements (CBMs) than either reading or early literacy measures, are given a detailed treatment in Chaps. 2–3 by some of the lead researchers in the field who are actively studying these measures in their respective labs. The chapters, focused on universal screening (Chap. 4) and response-to-intervention procedures (Chap. 5), take a broad stroke, applying those decisions across CBM academic areas and grade levels.

## Chapter 2

# Indicators of Fluent Writing in Beginning Writers

**Kristen D. Ritchey, Kristen L. McMaster, Stephanie Al Otaiba, Cynthia S. Puranik, Young-Suk Grace Kim, David C. Parker and Miriam Ortiz**

Learning to read and write are significant achievements in a child's education, and many children seem to pick up reading and writing skills with ease. They become fluent readers and writers and are able to coordinate reading and writing activities with seemingly little effort. But not all children learn to write with the same ease. Consider the case of Toby.

*Toby is a happy, precocious first grader who loves to learn. He listens intently when his teacher, Mrs. Wright, reads aloud to the class, and participates enthusiastically in class discussions. During independent reading time, he devours books about dinosaurs, fossils, and rocks. He loves science, and often draws intricate*

---

K. D. Ritchey (✉)  
University of Delaware, Newark NJ, USA  
e-mail: kritchey@udel.edu

K. L. McMaster  
University of Minnesota, Minneapolis MN, USA  
e-mail: mcmas004@umn.edu

S. Al Otaiba  
Southern Methodist University, Dallas TX, USA  
e-mail: salotaiba@smu.edu

C. S. Puranik  
University of Pittsburgh, Pittsburgh PA, USA  
e-mail: cpuranik@pitt.edu

Y. -S. G. Kim  
Florida State University, Tallahassee FL, USA  
e-mail: ykim5@fsu.edu

D. C. Parker  
ServeMinnesota, Minneapolis MN, USA  
e-mail: david@serveminnesota.org

M. Ortiz  
Southern Methodist University, Dallas TX, USA  
e-mail: mbortiz@mail.smu.edu

*pictures to illustrate the science topics discussed in class, describing them with detailed precision to his classmates. Mrs. Wright often jokes that he has the vocabulary of a paleontologist.*

*Despite his excitement about learning, Mrs. Wright has noticed that Toby has great difficulty with handwriting, spelling, and written composition. For example, Toby often shares elaborate stories of his family's camping excursions, relates what he recently learned at the science museum, or makes up harrowing adventure tales about himself and his dog. Yet, when Mrs. Wright encourages him to write about these ideas in his daily journal, he typically only writes a few words—such as “I like rocks,” “I went fishing,” “I saw a big bird.” Mrs. Wright has noticed that he forms each letter with painstaking care, and he often stops to ask her or a peer how to spell a word. Toby usually takes the entire 10 min allotted for journal writing to write these few words. Even then, his writing is often barely legible, with many erasures and crossings-out. Sometimes he gets frustrated and tears up his paper or wads it into a ball.*

*Mrs. Wright is concerned that, over time, Toby's struggles with writing will decrease his excitement about school and learning.*

This case example illustrates a child who is not a fluent writer, particularly due to his difficulties with handwriting and spelling, which comprise the transcription component of writing. Though Toby appears to have no trouble *generating* text—in the form of elaborate, articulate stories and explanations—getting his ideas onto paper is extremely difficult. His attention is consumed by forming letters and words, reserving little capacity for translating his sophisticated ideas into written text. For Toby, writing can be a frustrating activity, and his written production is not a good indicator of his ideas.

Early identification and intervention will be critical for students with writing problems, including students like Toby, to develop overall writing proficiency (Berninger, Nielsen, Abbott, Wijsman, & Raskind, 2008; Berninger et al., 2006), which will in turn have an impact on their long-term success in school and beyond (Graham & Perin, 2007). Early identification and intervention require assessments that can be used to establish current levels of writing proficiency, as well as to monitor progress in response to instruction and supplemental intervention. These assessments should focus on aspects of writing that serve as global indicators of fluent writing.

This chapter reviews existing work on the role of fluency in assessing writing, specifically when applying curriculum-based measurement (CBM; Deno, 1985) approaches to assessing fluent writing. We focus on work conducted at the early stages of writing development, from prekindergarten to third grade. This period is targeted because of the importance of early writing development, increased attention to writing instruction at these grades, and the need for further research that informs how fluent writing develops. We hope this chapter serves to further the conversation about appropriate approaches for assessing writing at early stages of writing development.

The first section of this chapter reviews the importance of writing and provides a description of writing fluency, drawing from definitions of reading fluency operationalized by CBM. It is followed by a review of measures that have been developed as indicators of writing; these measures directly or indirectly target fluency. Next, the chapter discusses the correlates of writing and how assessment can inform writing instruction and intervention. The chapter concludes with a discussion of directions for future research.

## Importance of Writing

Like reading, writing is important to academic and vocational success. By the time a child enters formal school settings, many foundational literacy skills have begun to develop, and the ability to express ideas in writing is further cultivated by academic instruction and experiences. The importance of writing has been highlighted in recent years for two reasons: the importance of writing to academic and vocational success and the poor writing of many students in the United States.

First, writing is critical to overall literacy development (Biancarosa & Snow, 2004). Writing provides students with the means to communicate what they know (Graham & Perin, 2007), is important for integrating knowledge and thinking critically (Shanahan, 2004), and can also “be a vehicle for improving reading” (Graham & Hebert, 2010, p. 6). Second, far too few students develop proficient writing. In fact, data from the most recent National Assessment of Educational Progress (National Center for Educational Statistics, 2012), indicate that only 30% of students in grades 8 and 12 performed at or above the “proficient” level (defined as solid academic performance) in writing. This widespread lack of writing proficiency is problematic given that a majority of jobs require employees to write proficiently at work (National Commission on Writing in America’s Schools and Colleges, 2004). In addition, many students with specific learning disabilities or other learning needs have more trouble writing than their peers do (Graham & Harris, 2005). For example, students with learning disabilities often experience significant difficulties with handwriting, spelling, and mechanics, as well as difficulties generating ideas and organizing, planning, and revising their writing (Troia, 2006).

These discussions have pointed to the need to improve writing instruction in schools. The National Commission on Writing (2003) called writing the “neglected ‘R.’” They called for comprehensive writing standards and proposed that writing should be built into “every curriculum area and at all grade levels ... from the earliest years through secondary school” (p. 5). While writing instruction has been identified as important and as an area for improvement at many points in the past decade (Biancarosa & Snow, 2004; Graham & Perin, 2007; National Commission on Writing, 2003), legislative attention has focused more heavily on improving reading, especially as part of *Reading First* and *No Child Left Behind* initiatives.

More recently, the development of the Common Core State Standards (National Governors Association & Council of Chief School Officers, 2010) has drawn increased attention to the instructional targets for writing for the US students. Nearly every state (45 states and the District of Columbia to date) has adopted the Common Core State Standards which delineate writing expectations for students beginning in kindergarten. The Common Core State Standards describe specific expectations for students to write across genres and to produce text that meets standards for writing conventions (e.g., syntax, mechanics). A summary of grade-level expectations is presented in Table 2.1. These expectations require high-quality instruction with sufficient opportunities to practice writing.

**Table 2.1** Common Core State Standards in Writing. (Adapted from the Common Core State Standards in Writing for K-third grade)

Writing Grades K-2		
Category		Skill
Text types and purposes	1.	Students should be able to create a product where they talk about a book and provide an opinion about the book with supported reasons.
	2.	Students should be able to create informative/explanatory texts where they are able to produce a topic and provide information on it.
	3.	Students should be able to describe or write a narrative event or sequence of events and discuss them with organization as well as react to what happened.
Production and description of writing	4.	Students should answer questions and respond to suggestions made by their peers in order to build up their writing.
	5.	Students should be able to participate in research projects, explore books, and express opinions in writing.
Research to build and present knowledge	6.	Students should be able to participate in research projects including exploring books and expressing opinions use them in writing.
	7.	Students should be able to remember information from personal experiences or be able to gather information from a provided source in order to answer a question.
Writing grade 3		
Category		Skill
Text types and purposes	1.	Students should be able to write opinion pieces with support for a point of view.
	a.	Students should be able to provide a topic, give an opinion, and organize reasons for that opinion.
	b.	Students should provide reasons for their opinion.
	c.	Students should use linking words and phrases in order to connect reasons and opinions.
	d.	Students should provide a conclusion statement or section in their writing.
	2.	Students should be able to write informative and explanatory pieces that examine a topic and conveys ideas and information clearly.

**Table 2.1** (continued)

Writing Grade 3		
	a.	Students should be able to pick a topic, examine it, and convey information and ideas about the topic clearly.
	b.	Students should be able to develop the topic using facts, definitions, and details.
	c.	Students should be able to use linking words and phrases to connect ideas and information.
	d.	Students should provide a conclusion statement or section in their writing.
	3.	Students should be able to write narratives to develop either real or imagined experiences, and include events while using effective techniques, descriptive details, and clear event sequences.
	a.	Students should be able to establish a situation, create a narrator and characters while organizing a sequence of events that has a natural progression.
	b.	Students should be able to use dialogue, describe actions, thoughts, and feelings of characters to create an experience or show the response of a character to a situation.
	c.	Students should be able to use temporal words and phrases to create organization and signal order.
	d.	Students should provide a sense of closure within their story.
Production and distribution of writing	4.	With scaffolding and guidance, students should be able to produce writing with clear organization and development for its intended purpose.
	5.	With scaffolding and guidance, students should be able to improve their writing as needed through the use of planning, revising, and editing.
	6.	With scaffolding and guidance, students should be able to use technology to publish their writing in addition to interacting and collaborating with others.
Research to build and present knowledge	7.	Students should be able to conduct research projects to facilitate building knowledge about a topic.
	8.	Students should be able to remember information from personal experience or be able to gather information from print sources as well as digital sources, and be able to sort the evidence into categories.
Range of writing	9.	Students should be able to write routinely over extended periods of times including time for research, reflection, and revision (or shorter time frames) for a range of tasks, purposes, and audiences.

Student performance on these standards will be assessed in newly developed assessment programs aligned to the Common Core State Standards. These assessments may require many school systems and teachers to use formative writing assessment and to provide feedback on their students' writing performance. With these new standards, it is possible that many schools will need to make substantial changes to



how they assess and teach writing in the early grades, especially as reading instruction has dominated literacy instructional time. It will be critical to consider stages of writing development within these assessments and instructional techniques. Even if high-stakes writing assessments are not included until later grades, the early grades should provide the foundation to develop critical writing skills including acquisition of and fluency in the component processes of writing. What remains unclear is the best way to accurately and efficiently assess writing proficiency and growth in a way that can inform instruction and maximize student progress within the curriculum and toward grade-level standards.

## **Fluent Writing**

Our current efforts have focused on the development of measures of writing for children in prekindergarten to third grade, with a specific interest in developing measures that fit into a CBM framework and that could serve as global indicators of fluent writing. The focus has been on defining appropriate tasks (i.e., what students are asked to do) and the scores (i.e., how the written product is evaluated) that provide the most technically adequate indices of writing proficiency.

### ***Using Principles of CBM to Define Fluency***

Work in CBM began almost 40 years ago by Stan Deno and colleagues as part of the Institute for Research on Learning Disabilities (IRLD) at the University of Minnesota. One of the goals of the IRLD was to develop a set of efficient and simple assessment procedures that could provide information about “vital signs” of academic health. These vital signs are often referred to as “global indicators” that can be used to indicate whether a student is making sufficient progress toward important academic goals, or whether a lack of progress indicates an underlying problem, such as a learning disability that requires further diagnosis and instructional changes or intervention (Deno, 1985).

Deno (1985) established key criteria for CBM. First, the measures are designed to provide information about a student’s proficiency and progress in academic areas such as reading, math, and writing. These measures use brief, direct observation of academic behaviors in ways that could be both efficient and yield scores that are reliable and valid indicators of academic outcomes. These academic outcomes are drawn from the curriculum to ensure alignment between the provided instruction and the assessment. Additionally, because the measures are to be used to monitor progress, they are designed to be sensitive to growth, meaning that the measures should yield scores that could be influenced by small amounts of learning. For example, the score might be expected to increase by one or two (or more) “points” when students were administered the tasks weekly or biweekly (Fuchs,

Fuchs, Hamlett, Walz, & German, 1996). Research on CBM has been conducted in core academic areas such as reading, mathematics, spelling, written expression, and content areas (see Foegen, Jiban, & Deno, 2007; McMaster & Espin, 2007; Wayman, Wallace, Wiley, Ticha, & Espin, 2007 for reviews).

CBM work in reading informs our work in developing CBM assessments for writing. For example, in reading, the most common CBM approach involves direct assessment of a student reading aloud from grade-level text while the number of words read correctly (and errors) in 1 min are recorded. At the elementary level, students' scores on CBM Passage Reading Fluency (also called Oral Reading Fluency) have strongly correlated with standardized reading measures (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Wayman et al., 2007), suggesting that CBM Passage Reading Fluency provides the type of global indicator of reading that Deno and colleagues were seeking. Further, the measures have been found to distinguish among students of different skill levels, to be useful for identifying students who may be at risk for reading disabilities (Jenkins, Hudson, & Johnson, 2007), and to be sensitive to growth made in brief time periods (Wayman et al., 2007).

We draw from the assertion that reading can be assessed by asking students to read text and that *reading fluency* can be identified as a global indicator of proficient reading. We posit that *writing fluency* can be assessed by asking students to write in response to a prompt and identifying the fluency of their response as a global indicator of proficient writing. We acknowledge that, as global indicators, this approach to assessment will not completely capture all of the many important aspects of writing, but that the derived scores do have instructional utility. Specifically, these indicators provide scores that can indicate whether a student is on track to meet important academic standards or is experiencing difficulties and is in need of further diagnosis and intervention.

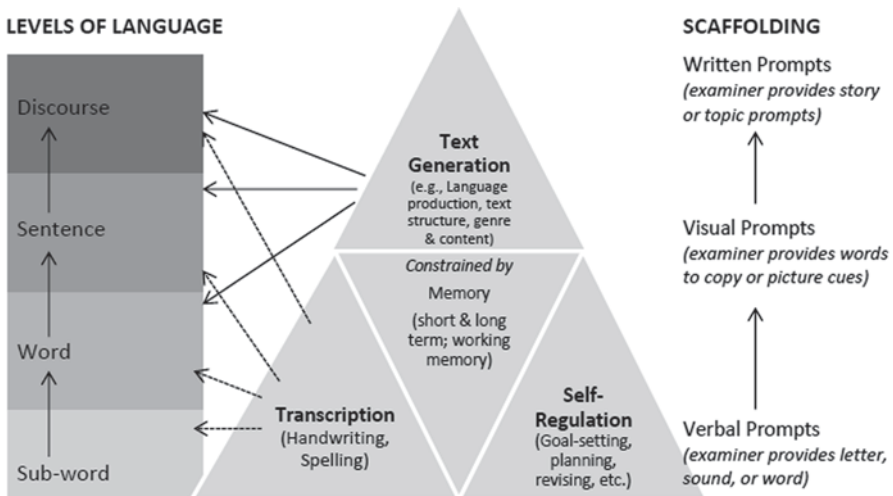
## ***Defining Fluent Writing***

Given our focus on developing global indicators of fluent writing, we turned to the CBM reading literature for a fluency definition. Deno and Marston (2006) define fluent *reading* as the way that “an individual easily processes text and that the processing of text encompasses both word recognition and comprehension” (pp. 179–180). Applying this definition to fluent *writing*, we propose that it is the way an individual easily *produces* written text, and that the generation of written text encompasses both *text generation* (translating ideas into words, sentences, paragraphs, and so on) and *transcription* (translating words, sentences, and higher levels of discourse into print). Thus, fluent writing comprises the ease with which an individual both generates and transcribes text. Below, we describe these components within a theoretical model of writing, provide operational definitions, and explain how the simultaneous execution and coordination of these components contributes to fluent writing.

## *Components of Fluent Writing: Transcription and Text Generation*

The text generation and transcription components of writing are derived from the seminal work of Hayes and Flower (1980), whose model of writing specified three key writing processes: planning, translating, and reviewing/revising. Researchers (e.g., Berninger, 2009; McCutchen, 2006) have further specified this model for early writing development. For example, Berninger and Amtmann (2003) described a “simple view of writing” that divides the translating process into two key components—text generation and transcription—and groups planning and reviewing/revising into a third component comprising self-regulatory processes. These three components can be presented in a triangle, such as that depicted in Fig. 2.1, with transcription and self-regulatory processes at the base, and text generation at the peak (Berninger & Amtmann, 2003). Fluent execution and coordination of these components is constrained by cognitive resources, such as short-term, long-term, and working memory (Berninger, 2009; McCutchen, 2006). Lack of automaticity (or execution of a process with little or no attention), in lower-level transcription processes constrains the higher-order processes involved in text generation, as well as for planning, organizing, and revising written text. For example, in the case example presented at the beginning of this chapter, Toby’s lack of automaticity in handwriting and spelling constrains his attentional capacity such that he is only able to generate and write down simple words and sentences even though his ideas are more complex.

Further, development of the transcription and text generation components of writing occurs at multiple levels of language, including sub-word, word, sentence, and discourse.



**Fig. 2.1** This figure illustrates the “simple view of writing” (Berninger & Amtmann, 2003). Fluent writing is constrained by memory, which can influence automatic execution of any of three components. Further, in the measures described in this chapter, transcription and/or text generation are assessed at each level of language. While self-regulation is not directly assessed, scaffolding is provided in the form of verbal, visual (e.g., pictorial), or written prompts to support children’s regulation of each type of task

and discourse levels (Whitaker, Berninger, Johnston, & Swanson, 1994). At the sub-word and word levels, children develop awareness of the alphabetic principle and graphophonemic relations and begin to transcribe letters, sounds, and words (Ehri, 1986). As children gain awareness and use writing conventions, they begin to separate words with spaces and thoughts with punctuation (Tolchinsky, 2006), and thus generate and transcribe text at the sentence level. As they gain knowledge of content and writing genres, they begin to produce longer units of writing at the discourse level (McCutchen, 2006).

Figure 2.1 illustrates how the component processes of writing (text generation, transcription, and self-regulation) are constrained by cognitive resources and develop across four levels of language (sub-word, word, sentence, and discourse). Work conducted by our research teams and others has included a search for global indicators of fluent writing for young children by tapping transcription (early in development) and text generation (as development progresses) across the four levels of language. Of note, self-regulatory processes in isolation have not been specific targets of these assessments; rather, the tasks provide varying levels of scaffolding (using verbal, visual, or written prompts) to support beginning writers' regulation of text generation. Below, we provide operational definitions of the components of the writing construct to be measured (transcription and text generation).

## Transcription

Transcription is the process of encoding sounds, words, sentences, and larger units of discourse into print, and involves both handwriting and spelling. For skilled writers, these skills are executed with automaticity, such that they require no or few attentional resources (e.g., Berninger, 2009; McCutchen, 2006; also see LaBerge & Samuels, 1974 for a seminal paper on automaticity). Handwriting involves the integration of orthographic coding [the "ability to represent a printed word in memory and then to access the whole word pattern, a single letter, or letter cluster in that representation" (Berninger & Rutberg, 1992, p. 260)] and those components of the motor system involved in executing the process of translating those words into print (Berninger, 2009). Beginning writers must allocate significant working memory resources to this orthographic-motor integration, which constrains higher-order writing processes. Handwriting is thus an important component of early writing assessment.

Spelling also involves orthographic coding, along with phonological coding (analysis and synthesis of phonemes in words; Berninger & Swanson, 1994). Like handwriting, spelling presents a significant challenge for young writers (e.g., Graham, Harris, & Fink-Chorzempa, 2002) and thus can place significant constraints on the development of other writing processes. Theoretical models of spelling development specify stages of qualitatively different approaches to spelling words (e.g., Ehri, 1986; Ehri & McCormick, 1998; Treiman & Bourassa, 2000a, 2000b). Accounting for differences across developmental stages in spelling is likely to be useful in capturing early indices of students' developing progress in writing.

## Text Generation

Text generation is the process of “turning ideas into words, sentences, and larger units of discourse” (McCutchen, 2006, p. 123), and is distinct from transcription of ideas into actual print (Berninger & Swanson, 1994). Text generation draws on linguistic sources including vocabulary knowledge (Coker, 2006; Kim et al., 2011; Olinghouse & Leird, 2009) as well as knowledge about topic and genre (McCutchen, 2006). As with transcription, text generation is constrained by cognitive resources. For example, working memory resources can constrain the writer’s ability to avoid grammatical errors and maintain linguistic connections within and across sentences and larger units of text. Long-term memory resources are related to knowledge of topic and genre, which can constrain quality and quantity of text generation.

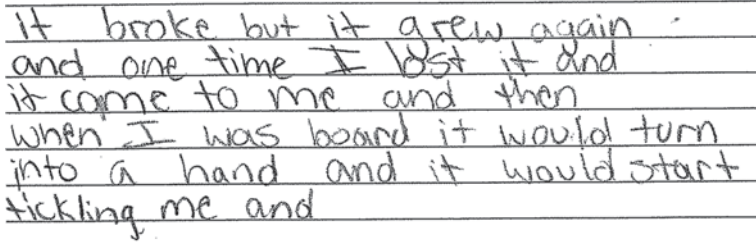
Like transcription, text generation has been demonstrated to be uniquely related to overall writing proficiency. Skilled writers are able to generate language more efficiently than less skilled writers, and this efficiency is a key predictor of writing quality (Dellerman, Coirer, & Marchand, 1996). This finding holds true for children just beginning to develop writing skills. For example, Juel, Griffith, and Gough (1986) reported that the number of ideas generated uniquely predicted first and second graders’ writing quality. Efficiency with language leads to greater language production and thus longer texts, and text length has been found to provide a strong index of text production as well as quality (Berninger & Swanson, 1994).

While transcription and text generation are distinct components that are predictive of overall writing proficiency, there is a necessary interplay between these components for writing to occur. Both transcription and text generation involve a complex coordination of component processes (e.g., orthographic, motoric, linguistic) that place considerable demands on cognitive resources needed for writing (Berninger, 1999). We hypothesize that measures that tap the development and automatization of transcription skills will serve as global indicators of children’s developing writing proficiency early on, but that measures that tap both text generation and transcription will quickly become important as children gain automaticity in their transcription skills. These global indicators are intended to identify students who may be at risk for writing difficulties and thus in need of further diagnostic assessment (beyond CBM), which would be used to develop specific interventions. In the next section, we describe assessments that hold promise as global indicators of beginning writing in prekindergarten to third grades.

## Assessment of Fluent Writing

Given our operational definition, our research teams have used the transcription and text generation constructs to develop and refine assessments that have potential to serve as indicators of fluent writing. As noted above, our work has been an extension of the work on CBM by Deno and colleagues which led to the development of the story prompt task (Deno, Mirkin, & Marston, 1982) which asks students to provide a written response for 3–5 min. An example of a story prompt completed by

**I once had a magic pencil and...**



it broke but it grew again  
and one time I lost it and  
it came to me and then  
when I was bored it would turn  
into a hand and it would start  
tickling me and

**Fig. 2.2** Story prompt response from a third-grade student

a third-grade student is presented in Fig. 2.2. The prompts are scored using production scores such as the number of words written, words spelled correctly, correct word sequences (which accounts for both spelling and grammar; Videen, Deno, & Marston, 1982), and correct minus incorrect word sequences (CIWS; Espin, Scierka, Skare, & Halverson, 1999). The definitions and an example of these scores are presented in Table 2.2.

Early research on CBM-writing tasks indicated that the measures were relatively simple and efficient to administer and score, and produced scores that were reliable and valid indicators of student writing proficiency for upper elementary students (Deno, Mirkin, & Marston, 1982; Parker, Tindal, & Hasbrouk, 1991; Tindal & Parker, 1991). One concern, however, was that the measures did not produce scores with technical adequacy for younger writers as strong as those produced for older writers. The measures yielded somewhat weak reliability ( $r = .20 - .47$ ; Deno, Mirkin, & Marston, 1982), and weak to moderate criterion validity ( $r = .23 - .63$ ; Gansle et al., 2004; Jewell & Malecki, 2005; Parker et al., 1991) for elementary students (as young as first grade). One possible explanation is that young students are unable to fluently produce sufficient text within the short time period or without some scaffolding of the writing activity, which could result in floor effects. This suggests that other tasks may be more appropriate for assessing the fluency of developing writers.

In the following sections, we describe tasks that could be useful as indicators of fluent writing for students in prekindergarten to third grade at the sub-word, word, sentence, and discourse levels of language. Some of these tasks were developed by our respective research teams, but we include the work of other researchers who have contributed to the current understanding of how to assess beginning fluent writing. We also include recent studies on using Story Prompt with first- to third-grade students. The tasks share the defining key criteria of CBM including alignment to curriculum and instruction, direct assessment of students, attending to a brief, standardized assessment time period, and using scoring procedures to index various aspects of writing proficiency which have the potential to monitor progress. Fluency, as operationally defined above, has been one of the aspects guiding the development of these measures.

**Table 2.2** *Scoring procedures*

There are several scoring procedures typically applied to index writing production. This text is a first-grade student's response to a Sentence Writing (Coker & Ritchey, 2010) prompt in which she was asked to write about an animal that lives on a farm. This was completed in the spring of first grade.

A photograph of a child's handwriting on a lined background. The text reads "A elyfint live on a farm". The word "A" is capitalized and the first letter of "elyfint" is lowercase. There are spaces between "A", "elyfint", "live", "on", "a", and "farm".

*Words written* scores are based upon the total number of words written, defined as any letter or group of letters separated by a space. Correct spelling and usage are not considered in the number of words written. In this example, there are six words written (a, elyfint, live, on, a, farm).

*Word spelled correctly* scores are based upon the number of words that are spelled correctly, given the context. In this example, there are three correctly spelled words (on, a, farm).

*Correct word sequence* scores take into account the correct spelling of each word, mechanics (capitalized word at the beginning of each sentence, terminal punctuation), and syntax of the response. A correct word sequence is marked with a  $\wedge$  and an incorrect word sequence is marked with a  $\vee$ . In this example, there are two correct word sequences. *Correct minus incorrect word sequence* is the difference score. In this case, there are five incorrect word sequences and the score is a negative number.

$$\vee A \vee \text{elyfint} \vee \text{live} \vee \text{on} \wedge a \wedge \text{farm} \vee$$

*Correct letter sequence* scores take into account the sequence of the letters within words. The first letter and the last letter in a word earn 1 and each correct sequence earns 1.

$$\wedge o \wedge n \wedge = 3 \text{ correct letter sequences}$$

$$\wedge f \wedge a \wedge r \wedge m \wedge = 5 \text{ correct letter sequences}$$

$$\wedge e \wedge l \vee y \vee f \vee i \vee n \wedge t \wedge = 4 \text{ correct letter sequences}$$

Table 2.3 lists the assessments and the proposed grade levels when they may be useful as indicators of fluent writing. In the early grades (prekindergarten and kindergarten), the curriculum focus is on writing names, letters, and words. For students in kindergarten and first grade, the curriculum focus is on spelling words and writing text, such as sentences, and these foci continue and expand for second and third grades. Both transcription and text generation skills are included in aspects of the assessments. In the following sections, we describe assessments that have the potential to serve as indicators of fluent writing at the sub-word, word, sentence, and discourse levels of language.

## *Name Writing*

At the prekindergarten level, the development of writing is nascent, and the development of emergent literacy skills and fine motor skills that support writing are often the focus. Drawing and simple writing are the types of activities in which children might engage, especially in earlier months of prekindergarten (Tolchinsky, 2006). Some early writing skills that children demonstrate include writing their names and letters of the alphabet. Writing may also be supported by a teacher who



**Table 2.3** Assessments by grade level

Pre-K	Kindergarten			First Grade			Second Grade			Third Grade			
	All Year	Fall	Winter	Spring	Fall	Winter	Spring	Fall	Winter	Spring	Fall	Winter	Spring
← Name Writing →													
← Letter Writing →													
← Letter Copying →													
← Word Dictation →													
← Spelling →						← Spelling →							← Spelling →
						← Sentence Writing →							
						← Sentence Dictation →							
						← Picture-Word →							
										← Picture Story Writing →			
										← Picture Prompt →			
										← Story Prompt (CBM-W) →			



transcribes a student's ideas or students might dictate "stories" to adults, who then write them down.

One writing indicator that could represent a global indicator of early writing at this level is name writing. Knowing how to write one's name appears to be an important literacy achievement; it is often the first word children learn to write because of its social and personal significance to the child. When children first start writing, they use the first letters of their names to spell the other words (e.g., Bothde Vries & Bus, 2008, 2010; Treiman, Kessler, & Bourassa, 2001) perhaps serving as a model for future writing (Bloodgood, 1999; Ferreiro & Teberosky, 1982). Past research has highlighted the important role of name writing in the development of children's spelling skills. For example, the National Early Literacy Panel (Lonigan, Schatschneider, & Westberg, 2008) reported a moderate relation of  $r = .36$  between children's name writing and spelling. Other research has shown that children with more advanced name writing skills spell more words (Bloodgood, 1999; Bothde Vries & Bus, 2010; Levin et al., 2005; Puranik & Lonigan, 2012; Treiman & Broderick, 1998; Welsh, Sullivan, & Justice, 2003) leading some to suggest that name writing could be used as a screener for children's literacy skills (Haney, 2002; Haney, Bissonnette, & Behnken, 2003).

Assessing name writing skills can occur formally or informally, whereby children write their names in authentic situations (e.g., drawing or signing their names on greeting cards). An alternative to counting the number of letters or percentage of letters is the use of a rubric scoring system. Several rubrics have been proposed to examine name writing (Bloodgood, 1999; Diamond, Gerde, & Powell, 2008; Haney et al., 2003; Levin & Bus, 2003; Molfese et al., 2011; Puranik & Lonigan, 2011; Sulzby et al., 1989; Welsh, Sullivan, & Justice, 2003). To evaluate whether any one of these rubrics was a better indicator of children's name writing skills, Puranik, Schreiber, Estabrook, and O'Donnell (2014) compared children's name writing scores on six rubrics including a simple scale and examined their correlation to literacy skills (letter names, letter sounds, phonological awareness, print concepts, and spelling) in a sample of 346 preschool children pooled across four studies. Scores from these rubrics were highly correlated ( $r = .62-.94$ ). Further, the magnitude of the correlations among name-writing scores and children's literacy skills were similar ( $r = .52-.60$  for letter writing,  $r = .30-.39$  for phonological awareness,  $r = .38-.49$  for letter sounds,  $r = .46-.54$  for spelling). This suggests that these rubrics provide similar indices of children's name writing skills at a single assessment point.

Despite the importance of name writing, it has limited use as an indicator of early writing development. For example, researchers have noted incongruities in children's knowledge regarding the letters in their names, and that children often write their names by rote. These observations have led researchers to contend that name writing reflects procedural, rather than conceptual knowledge of letters (Drouin & Harmon, 2009; Puranik & Lonigan, 2012). Another limitation is that, by the end of preschool, some children appear to be able to write their names conventionally and their scores (across multiple studies) show ceiling effects (e.g., Puranik & Lonigan, 2012). In summary, name writing appears to be better suited for predicting literacy

skills early in the preschool years, but may not be robust for tracking children's progress in acquiring and developing fluent writing over time.

To date, researchers have not focused specifically on fluent name writing, yet fluency is likely an important aspect of name writing. Most teachers have worked with at least one student who slowly and laboriously writes his or her name on school papers, indicating dysfluent name writing. How to capture fluent name writing is a measurement challenge. While timing a child's writing is one way to capture fluency, it may not generalize to a child's name or provide a meaningful score as names can vary in length and in the difficulty of the letters in any given name. Given the limitations with name writing noted above, and difficulties related to assessing fluent name writing, other approaches to assessing development of fluent writing are likely needed.

### *Letter Writing Assessments*

By the end of the preschool years, children's letter writing may be a better indicator of developing writing skills than their name writing skill. Assessments of letter writing have included copying, production (untimed), and fluent production (timed) measures. For example, VanDerHyden, Witt, Naquin, and Noell (2001) studied a letter-copying task that was part of a readiness battery with kindergarten students. Students were provided with a sheet of paper with uppercase letters, and instruction to copy the letter in a space below each letter; the score was the number of correct letters copied in 1 min. Alternate-form reliability for two forms (one with letters in ascending order and one with letters in descending order) was  $r = .68$  ( $N = 107$ ). Validity coefficients of Letter Copying with the Onset Recognition Fluency subtest of Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 1996), and the Comprehensive Inventory of Basic Skills (Brigance, 1999) ranged from  $r = .21$  to  $.59$  ( $N = 40$ ). These coefficients do not provide evidence of strong reliability and validity, but the order of letters within the alternate forms was noted by the authors as possibly introducing measurement error, and the measures were administered in the second half of kindergarten in a group-administration format. It is possible that this type of task could be a better index for younger children's writing, such as in prekindergarten or early kindergarten, but the recommendation requires additional research.

Untimed letter production tasks, such as Letter Writing (Ritchey, 2006), are designed to capture transcription skills and students' ability to reproduce letters from memory (i.e., without a model). In this assessment, letters of the alphabet are dictated to students and they are asked to write both the upper and lowercase letter provided by the examiner. Subsequent letters are dictated when the student is finished or indicates "I don't know." In pilot work, letters were administered at a set interval (every 10 s, similar to spelling CBM administration procedures described below), but it was determined to be difficult for many students who were unable to quickly retrieve the letter form and write the letters, overly time consuming for

students who wrote letters quickly, and frustrating for students who identified an error, but did not have time to make corrections. Responses are scored based on whether or not the letter could be identified in isolation. Alternate forms of Letter Writing were created by randomly selecting the order of letters. Split-half reliability coefficients for letter writing was in the .90 range, suggesting appropriate internal consistency of all forms. The concurrent criterion-related validity coefficients were .61 with the Test of Early Reading Ability, 3rd Edition (TERA-3; Reid, Hresko, & Hammill, 2001), and .72 with Letter Name Fluency.

Two limitations of an untimed measure include the possibility of ceiling effects (especially at later stages of letter writing development) and the inability to capture individual differences in fluent production. One alternative is to determine the total amount of time required to complete all letters and calculating a rate score (number of letters per minute). Another approach would be to further investigate dictating letters at a set interval, which would also allow for standardization for small-group or whole-class administration. Others have investigated letter writing fluency tasks, described below.

Letter writing fluency—the ability to retrieve and write the letters of the alphabet as quickly as possible under timed conditions—has been repeatedly shown to be an excellent indicator of both writing quality and quantity in older elementary school children (e.g., Graham, Berninger, Abbott, Abbott, & Whitaker, 1997; Jones & Christensen, 1999). Berninger and Rutberg (1992) found a strong correlation between the letter fluency task and spelling and composition in their study with first-, second-, and third-grade students and concluded that the letter fluency task as measured in 15 s has adequate concurrent validity for assessing beginning writing.

Puranik, Al Otaiba, Folsom, and Greulich (2013) recently examined Letter Writing Fluency in a sample of 102 kindergarten children. Participants were recruited from eight classes in four different schools across three districts. The mean age of the participants was 69.6 months ( $SD=4.3$ ), and gender was equally distributed (50 males and 52 females). Letter Writing Fluency was administered in a whole-class format and children were asked to write the lowercase letters of the alphabet as quickly as possible in 1 min. In addition, these children were also individually administered standardized tests of spelling and writing. Assessments were completed at the end of the fall.

The mean score for Letter Writing Fluency was 4.72 letters per minute ( $SD=4.06$ ; Range 0–23), although there was large variability in the number of letters children were able to write. The score was positively related to DIBELS Letter Naming Fluency ( $r=.58$ ). Criterion-related validity was investigated using the Test of Early Written Language-3 (TEWL-3; Hresko, Herron, Peak, & Hicks, 2012) and the Spelling subtest of the Woodcock–Johnson Tests of Achievement, Third Edition (WJ-III; Woodcock, McGrew, & Mather, 2001, 2007). The validity coefficients were  $r=.56$  and  $r=.44$  respectively. Thus, preliminary evidence indicates that Letter Writing Fluency may present a potential option for assessing writing fluency in kindergarten children.

An important consideration in assessing letter writing as an indicator of fluent writing of young children is that all letters are not of the same difficulty, both in

letter formation and in children's exposure, and opportunities to practice them. Recently, Puranik, Petscher, and Lonigan (2012) examined the dimensionality and reliability of letter writing skills in 471 preschool children with the aim of determining whether a sequence existed in how children learn to write the letters of the alphabet. Puranik et al.'s findings indicate that the ability to write some letters is acquired earlier than the ability to write other letters. Furthermore, results of item response theory analyses indicated the existence of an approximate developmental sequence to how children learn to write uppercase letters. The 10 easiest letters for preschool children to write were *O, L, A, B, X, T, H, I, E,* and *P*, whereas the 10 hardest letters to write were *J, K, Z, G, Q, V, U, Y, R,* and *N*. In contrast, there was less evidence for a clear sequence for letters in the middle range of difficulty (e.g., *C, D, F, S,* and *W*). These findings could be considered when assessing letter writing. For example, if the goal is to assess school readiness, letters with high discrimination but low difficulty (e.g., *L, I, T, X,* and *E*) could be included in the assessment; however, to distinguish among more precocious students, letters with high discrimination and high difficulty (e.g., *G, K, R, U, V,* and *Y*) may be more appropriate. These findings could aid in monitoring children's progress of developing letter-writing skills.

### ***Word Writing and Spelling***

Transcribing, spelling, and generating words are important to writing development. It is possible that tasks involving such transcription processes at the word level could serve as a global indicator of beginning writers' overall writing proficiency. These tasks may have the greatest utility for students in kindergarten and first grade, or as indices of spelling. Word-level measures such as copying, dictation, and spelling target the transcription component of writing, and word generation tasks also target the text generation component of writing.

Lembke, Deno, and Hall (2003) investigated a copying task that fits into the framework as a word-level transcription assessment. In Word Copying, students are presented with written words and are asked to copy each word on lines below the word, and the score relates to the amount of text produced in 2 min. Scores include correct letter sequences, number of words, and words spelled correctly. The criterion measure was an "atomistic" score from a story prompt writing sample (average words written, words spelled correctly, correct word sequences, and correct minus incorrect word sequences). The validity coefficients ranged from  $r = .06$  to  $.69$  for a sample of 15 students enrolled in a summer program (rising second grade). Twelve of the 24 coefficients were at or exceeded  $.50$ . These values may be relatively low, given the developmental level and small sample of students.

In a second study examining Word Copying, Hampton, Lembke, and Summers (2013) investigated several characteristic of the measure, including the effect of manipulating the amount of time that students were provided to write and the effect on reliability and validity. A 1, 2, and 3 min score (using correct word sequences and correct minus incorrect word sequences) was calculated. The alternate-form

reliability for 1 min ranged from .34 to .72, for 2 min ranged from .56 to .86, and for 3 min ranged from .71 to .98. The 3 min score generally yielded stronger coefficients. Using the Test of Early Written Language, 2nd Edition (TEWL-2; Hresko, Herron, & Peak, 1996) as the criterion measure and the correct word sequence score, the validity coefficient for 1 min was .39, for 2 min was .30, and for 3 min was .44. Similar to reliability, the 3 min score generally yielded the strongest validity coefficients. These tasks show promise as a way to assess fluency with transcription tasks, but may be most appropriate for younger students. This task has not, however, been studied with younger students.

Other tasks at the word level of language require students to generate the spelling of the word. Spelling assessments vary in the type of words that children are asked to spell (regular or irregular spellings, real words or nonsense words) and the manner in which the assessments is administered (timed, untimed, delivery of words at set intervals, or timing of completion of the whole task). Spelling tasks can assess word-level transcription, using decodable and high-frequency-word spelling words. For example, Spelling Fluency includes words that are randomly drawn from the Harris–Jacobson grade-level list (Deno, Mirkin, Lowry, & Kuehnle, 1980; cited in Fuchs, Fuchs, Hamlett, & Allinder, 1991a, 1991b). Students are asked to spell a new word every 10 s for 2 min. The scores calculated include correct words and correct letter sequences per minute. Tindal, German and Deno (1983, cited in Shinn & Shinn, 2002) reported alternate-form reliability ranging from .86 to .97 (for grades 1 to 6). Deno, Mirkin, Lowry, & Kuehel, (1980) reported strong test-retest reliability of .92, and criterion validity based on the Spelling subtest of the Peabody Individual Achievement Test and the Spelling subtest of the Stanford Achievement Test ( $r = .73$  and  $.99$ ), respectively for second- to sixth-grade students.

This measure is similar to the CBM spelling materials available from AIMSweb (Shinn & Shinn, 2002). Spelling words are randomly selected from grade-level sets of words taken from words that were found across available literacy curriculum. There are 20 alternate forms of first-grade probes, and 30 alternate forms of second- and third-grade probes. One important difference between the spelling measures is the interval of word delivery by the examiner. For students in first and second grade, a new word is administered every 10 s and for students in third grade and above, a new word is administered every 7 s. Responses are scored for correct words and correct letter sequences.

Another task, Word Dictation (Lembke et al., 2003) adopted the procedures for developing CBM spelling tasks, but for this version sets of spelling words were identified for each alternate form by sampling words from the Basic Elementary Reading Vocabulary list (Harris & Jacobson, 1972). To develop alternative forms, words are randomly selected and administration is timed (3 min). The administration of this task differs from CBM Spelling. In this task, words are dictated when students have finished writing the previous word or after a 5 s pause. Each word is repeated twice and responses are scored for correct letter sequences, words written, and correct words.

Word Dictation was also investigated in the Hampton et al. (2013) study, and the time interval was manipulated. The words for this version of the measure were selected from the Fry 100 Words List (Fry, Kress, & Fountoukidis, 1993) that had seven or fewer letters. The alternate-form reliability for 1 min ranged from  $r = .55$  to  $.92$ , for 2 min ranged from  $r = .66$  to  $.92$ , and for 3 min ranged from  $r = .58$  to  $.97$ . The validity coefficients for this measure (also using TEWL-2 as the criterion measures) were  $r = .50$  for 1 min and  $r = .48$  for both 2 and 3 min. One important difference between the two tasks is the time structure. With a standardized, set timing of Word Dictation, it is possible that the assessment could be group administered, but approaches that provide subsequent words to students as they finish may be most appropriate for individual administration. These are important considerations if spelling or other word-level tasks are to be efficiently administered.

The ability to spell a word provided by an examiner or teacher is one aspect of word-level writing skills. However, another aspect of writing competence is the ability to generate words and transcribe them efficiently. McMaster, Du, and Petursdottir (2009) developed a task that required students to generate, as well as to spell, words using letter prompts. Three alternate forms of Letter Prompt were administered to 50 first graders. Each form contained four pages with 54 letters total (with repetition). The letters were selected randomly from all letters in the alphabet except *q*, *x*, *y*, and *z*, which are used infrequently by first-grade students (these letters were used for practice). After practicing with the examiner, students wrote as many words as they could that started with the letters provided. After 3 and 5 min, students circled the last letter they had written. Letter Prompt was scored for words written, words spelled correctly, and correct letter sequences. Alternate-form reliability ranged from  $r = .38$  to  $.81$ , with stronger reliability for 5 min samples. An important consideration is the grade level; this measure may be a better indicator for younger writers.

The preliminary efforts in developing CBM measures for young children suggest that assessing sub-word and word-level components writing may serve as global indicators of fluent early writing. In turn, future research is needed to test whether these components are indeed good predictors of longitudinal growth, whether the skills they assess are malleable to early intervention, and whether the measures themselves are sensitive enough to growth to be used diagnostically in a RTI framework. These tasks focus heavily on transcription skills, which align to one component of the theoretical model. Several tasks (such as Word Copying or Spelling) may be applicable for students who are younger than those originally studied, which may be an avenue for additional research. For example, Word Copying or other word-level tasks (spelling simple words, such as *c-v-c* words) may be tasks appropriate for prekindergarten or kindergarten and serve to fill in a gap in assessments for young writers.



## *Sentence Writing*

Several measures have been developed that focus on fluent writing of sentences, and these measures serve to bridge the word-level writing proficiency to extended discourse-level writing proficiency. For young children, especially in kindergarten and beginning first grade, measures that assess only word-level writing proficiency may not be challenging enough (e.g., spelling *c-v-c* words). An assessment that asks students to write a sentence or several sentences may be one way to elicit extended responses, but in a way in which there are modest expectations for the length of the text produced, the topical complexity (such as the number of details), and the type of planning required. Assessments at the sentence level have included measures that focused on transcription only or included both transcription and text generation. These measures have typically been conceptualized for students in later kindergarten and first grade, as they may have limited applicability for students with more fluent writing proficiency.

One measure, Sentence Copying (Lembke et al., 2003, McMaster et al., 2009, McMaster, Du, Yeo, Deno, Parker, & Ellis, 2011), is a sentence-level measure designed to focus only on transcription. In each alternate form, students are presented with 15 sentences of 5–7 words in length, and each sentence has a lined space under each sentence on which the students are to copy the sentence. Sentences are selected from existing language arts curriculum materials and include simple statements such as “*We have one cat.*” and questions such as “*Who is he?*” After a practice item, students copy the remaining sentences in the allotted time (3 and 5 min time periods were studied). Sentences are scored for words written, words spelled correctly, correct word sequences, and correct minus incorrect word sequences.

Lembke et al. (2003) also investigated a sentence writing measure, and they reported a wide range of validity coefficients, with as low as  $-.01$  and as high as  $.81$  (19 of the 32 coefficients exceeded  $.50$ ) using a writing sample as the criterion measure for rising second-grade students using a 3 min administration. The criterion measure was a writing sample evaluated using atomistic scores (described above). McMaster et al. (2009) compared scoring options and administration time (3 vs. 5 min) and reported reliability coefficients ranging from  $r = .71$  to  $.85$  for 3 min and  $r = .76$  to  $.89$  for 5 min for first-grade students. Validity coefficients ranged from  $r = .48$  to  $.67$  for 3 min and  $r = .43$  to  $.67$  for 5 min using a teacher rating as the criterion measure;  $r = .67$  to  $.70$  for 3 min and  $r = .53$  to  $.65$  for 5 min using a writing sample scored using a district writing rubric as the criterion; and  $r = .33$  to  $.44$  for 3 min and  $r = .32$  to  $.47$  for 5 min using the Test of Written Language, 3rd Edition (TOWL-3, Hammill & Larsen, 1996) as the criterion measure. Overall, Sentence Copying appears to have reasonable technical characteristics. However, one important limitation include that many teachers view copying as having low-face validity, that is, they do not view copying as an end goal of writing instruction which could limit adoption.

In the Lembke et al. (2003) study, Sentence Copying was compared to Sentence Dictation, another task that includes transcription skill, but required students to

spell words in the sentences and attend to mechanics such as capitalization and punctuation, rather than only copy the presented sentences. Text generation was not a construct assessed in this measure. For Sentence Dictation, sentences that ranged in length from 5–7 sentences are dictated to students two times, and students are provided with the next sentence after a 5 s pause or completion of the sentence. Sentences are administered for 3 min each, and the total amount of text produced by the student in that time period is determined. Responses were scored for words written, words spelled correctly, correct word sequences, and correct minus incorrect word sequences. Validity coefficients ranged from  $r = .27$  to  $.92$ , with 28 of the 32 coefficients greater than  $.50$ . The criterion measure was an “atomistic” score described above. In general, the technical adequacy of Sentence Dictation was stronger than Sentence Copying for this grade level when atomistic ratings of writing were evaluated. It also served to address concerns expressed by teachers who believed that copying tasks were not valid indicators of writing. However, copying may be an indicator of how well a writer can coordinate the orthographic-motor processes required in writing and appropriate for younger students as an indicator of fluent writing.

Two other sentence writing assessments have been developed, and include attention to fluent transcription and text generation. Sentence Writing (Coker & Ritchey, 2010) was developed as a measure that could capture both text generation and transcription. In this task, students are given a topic and asked to generate text. In this task, two sentence prompts are orally presented. (e.g., *Chocolate chip cookies are yummy. Write about your favorite cookie.*). After each prompt, students are given up to 3 min to write. Student can respond by writing a word, phrase, or sentence in response to the sentence prompt. Students’ written production is evaluated using words written, words spelled correctly, and correct word sequences. (A qualitative score was also developed and is discussed below). Alternate-form reliability for Sentence Writing was  $r = .74$  and  $.77$  for words written,  $r = .80$  and  $.75$  for correct word sequences, and  $r = .81$  and  $.87$  for words spelled correctly for students in spring of kindergarten and first grades. Using Test of Early Written Language; 2nd Edition (TEWL-2; Hresko, Herron, & Peak, 1996) for kindergarten as the criterion variable, and WJ-III Spelling, Writing Samples, and Broad Writing composite scores (WJ-III; Woodcock et al., 2001, 2007) as the criterion measures for first grade, the validity coefficients ranged from  $r = .20$  to  $.46$  in kindergarten and  $r = .25$  to  $.57$  for first grade.

A final measure of sentence-level writing that includes attention to both transcription and idea generation is called Picture-Word (McMaster et al., 2009; McMaster, Ritchey, & Lemke, 2011). Fluent writing is assessed via students’ responses to a picture paired with a word that describes the picture (e.g., a picture of a hat with the word *hat* written below it). Providing a picture and a word for students was hypothesized to provide support and scaffolding for first-grade students who were beginning writers. Students are given 3 min (after a practice item) to write, and there are multiple pages of picture-words so that students will not finish the entire task in the allotted time. Students’ writing production is scored for the words written, words spelled correctly, and correct word sequences. Reliability estimates ranged from



$r=.44$  to  $.79$  for 3 min and  $r=.53$  to  $.79$  for 5 min, and validity estimates ranged from  $r=.37$  to  $.60$  for teacher rating and writing sample scored using a district rubric, but were not statistically significant for the TOWL-3.

In summary, several assessments have focused on fluent sentence writing and have included copying, dictation, and composition formats. The reliability of the measures varies, often depending on which score is being considered. A clear pattern of which score (or scores) is the most appropriate index of sentence-level writing proficiency is not evident. For some tasks, words written yields highest reliability estimates, but other scores yield higher reliability for other assessments. Validity estimates vary widely, and these differences could be attributed to different criterion measures used across studies (i.e., different norm-referenced assessments, teacher rating, and writing samples scored with divergent approaches).

One challenge for assessments that require text generation (both for sentences and discourse-level tasks) is that the writer produces the text that is evaluated, not the assessment. Consider a spelling assessment. The examiner provides the word, and the student is asked to reproduce that word; scoring is based on the degree to which the student's response matches the correct spelling. In an assessment that requires text generation, students can produce text that varies widely. In the sentence writing assessment described above, first-grade students responded to the prompt about animals that lived on a farm with widely varied responses, including writing about only one animal such as a goat or a horse, writing about multiple animals that could live on a farm, including common pets such as dogs and cats, and a response about an elephant (which though unlikely, is not impossible). Students could select words that were easier to spell (e.g., *pig* instead of *alpaca*), use a strategy for spelling that uses words found in the printed prompt for support in spelling (e.g., use the spelling of *animals* and *farm* as models), or use invented spelling for words that were unfamiliar. The spelling of words may be legible to the scorer, but the child's spelling of some words can be so unclear that a scorer is unable to identify the intended word or meaning. In scoring writing samples, there are no correct or incorrect criteria for children's wide range of ideas, which can challenge the development and implementation of any scoring system. Thus, the state of the art of assessment to inform instruction and intervention is evolving. Converging evidence supports the promise of writing CBM for diagnostic purposes, but there are not yet a set of evidence-based practices for assessing progress or for conceptualizing benchmarks. These scoring challenges extend to discourse-level writing measures, which include those described in the following section.

## ***Discourse Writing***

A final set of assessments that could be used as indicators of fluent writing are discourse-level measures. These measures ask children to write extended text, typically on a single topic. These assessments have the greatest potential to be used as global indicators across grade levels, as the ability to write extended discourse with

fluency is an important goal for skilled writers. The procedures and prompts could be applied both within and across grade levels. Fluent writers are able to plan and organize ideas quickly and efficiently as well as revise and edit text. For younger children, discourse-level writing abilities may be limited to short, simple narratives or descriptions. An important consideration in developing discourse-level tasks for beginning writers is that text generation may be constrained by transcription skills. Students may be unable to produce large amounts of text, and students may get “stuck” trying to spell words or transcribing ideas as quickly as they can generate ideas. In the following sections, we describe several measures that have been investigated with younger writers.

As described above, the first CBM measure of written expression was the CBM Story Prompt designed by Deno and colleagues. The task was originally proposed for first to sixth grades (Deno, Mirkin, & Marston, 1982; Marston & Deno, 1981), and has been evaluated across these grade levels and extended upward to 11th grade (see McMaster & Espin, 2007, for a review). In contrast to CBM for reading and math, Story Prompt is the only CBM measure with assessment probes that are not grade-specific, and it can be used both within and across grades. In this task, students are provided with a story starter such as *One day, I was walking home from school and ...* or *It was a cold and rainy day. All of a sudden ...* and were asked to finish the story. Sets of writing prompts are available from AIMSweb (<http://www.aimsweb.com>) and the Research Institute on Progress Monitoring ([www.progressmonitoring.org](http://www.progressmonitoring.org)). Each alternate form of Story Prompt includes a sentence or sentences that begin a story, a set of ellipses, and lined writing space for the student to write the story. Students are provided with 1 min to plan and 3 or 5 min to write (which can vary by developer or grade level). Scores are based on the amount of text that students can produce in the given time period and include counting words written, correctly spelled words, correct letter sequences, correct word sequences, and/or correct minus incorrect word sequences. McMaster and Espin (2007) summarized 11 studies (out of 28 reports and published articles) that included first-, second- and third-grade students and reported a wide range of reliability ( $r = .006-.96$ ) and validity ( $r = -.24-.88$ ) using scores for words written, words spelled correctly, percentage of words spelled correctly, correct word sequences, percentage of correct word sequences, correct minus incorrect word sequences, and T-units, an index of mature and large words.

More recent work has studied this assessment in early elementary grades (McMaster et al., 2009; Ritchey & Coker, 2013). Some of this work has taken a comparative approach by studying whether a novel measure (or measures) yields scores that improve the reliability and validity of scores as compared to those of Story Prompt. For example, McMaster et al. (2009) used Story Prompt and evaluated whether 3 or 5 min of writing time was most appropriate in yielding reliable and valid scores. After 3 min, the examiner instructed participants to stop and raise their pencils in the air, circle the last letter they had written, and then continue writing until a total of 5 min had passed. Students wrote in response to a single prompt, but the text produced in each time interval was evaluated. Test-retest reliability ranged from  $r = .49$  to  $.74$  for 3 min and  $r = .45$  to  $.83$  for 5 min. Alternate-form reliability

was similar, and also slightly higher for 5 min ( $r = .47-.75$  for 3 min;  $r = .54-.83$  for 5 min). Ritchey and Coker (2013) investigated the validity of Story Prompt, using the WJ-III Writing Samples (Woodcock et al., 2001, 2007) and a teacher rating (1–5 scale) as the criterion measures. The validity coefficients with WJ3 Writing Samples ranged from  $r = .31$  to  $.42$  and with teacher rating ranged from  $r = .30$  to  $.48$  for words written, words spelled correctly, and correct word sequences.

Several other measures have added a picture to the writing prompt. Picture Story (Ritchey & Coker, 2013) is a discourse-level writing task for first to third grades. Students are asked to write a story that matches the events portrayed in a sequence of three pictures. To date, three sets of prompts have been evaluated by Ritchey and Coker. The prompts include three pictures (e.g., a picture of a dog covered with mud, a picture of a dog getting a bath, and a picture of a clean dog) and students are asked to write for 5 min. Production scores (including words written, words spelled correctly, and correct word sequences) have been evaluated. The alternate-form reliability ranged from  $r = .78$  to  $.83$ ; validity coefficients ranged from  $r = .37$  to  $.42$  with WJ-III Writing Samples (Woodcock et al., 2001, 2007) and  $r = .29$  to  $.50$  with a teacher rating.

The tasks described above have primarily been focused on narrative writing, which is an important writing skill, but it is not the only writing skill and curriculum focus. These other genres become increasingly more prevalent in the curriculum and writing expectations as students move to higher grades, including broader coverage in the Common Core State Standards. McMaster and Campbell (2008) examined whether different types of discourse-level measures (e.g., narrative, expository, and photo writing prompts) that were previously validated with older students in fifth and seventh grades could serve as indicator for third-grade students. They found that photo prompts (which prompted students to write about a photograph of children doing school-related activities) administered for 3 min yielded sufficient reliability ( $r > .70$ ) and validity ( $r > .50$ ) coefficients within third grade, and that 5 min narrative story prompts administered for 5 min yielded sufficient reliability and validity coefficients.

In describing one final assessment, we offer an example of a task designed for discourse-level writing that was deemed not appropriate for beginning writers because of a possible misalignment of the complexity of the writing demand and the writing abilities of first-grade students. McMaster et al. (2009) extended three discourse-level tasks to first grade. These tasks included Story Prompts and Photo Prompts (described above), and a third type of prompt: Picture-Theme. Picture-Theme included three prompts with the following themes that were identified (with participating teachers' help) as familiar to most first graders attending the US public schools in the Midwest: birthday party, snow, and school. These prompts used high-frequency words from the literacy curriculum (Houghton-Mifflin). Each theme contained four pictures with one related word underneath each picture. For example, the "birthday party" theme had the pictures of cake, friends, home, and birthday balloon with the four words: *cake*, *friends*, *home*, and *birthday* printed underneath each picture. The pictures came from Microsoft clip art and researcher-drawn pictures. Each theme prompt was printed at the top of the first page with lines printed below

the prompt. Additional lined sheets were provided in case they were needed. Before students were asked to write, the examiner first instructed the participants to identify the four pictures provided and then asked the participants to write a story based on the theme. Picture-Theme was scored for words written, words spelled correctly, and correct word sequences, and correct letter sequences for 3 and 5 min writing time. None of these scoring procedures produced particularly strong alternate-form reliability ( $r = .31-.70$ , with most coefficients below  $.70$ ). Criterion-related validity coefficients with teacher ratings and a district scoring rubric ranged from  $r = .37$  to  $.60$ . Given relatively weak reliability coefficients compared to other sentence-level (Picture-word) and discourse-level (Story Prompt) measures, Picture-Theme was not viewed as promising indicator (McMaster et al., 2009).

We offer this example to illustrate the challenge of identifying appropriate tasks and appropriate supports in measures designed as fluency-based global indicators for beginning writers. Specifically for Picture-Theme, we speculate that the writing demands and the cognitive resources required to use a specific set of words and to write on a specific topic may have required more planning and organization on a student's part than a more open-ended prompt. Another possible explanation is that perhaps 5 min was not enough time for students to plan, organize, and transcribe ideas, and a longer time period may be necessary to yield sufficient text to evaluate. These aspects of the assessment may, in part, explain the weaker reliability and validity coefficients. Another concern with the use of CBM at early stages of writing development is the complexity of the task and determining the appropriate writing topic for students. The topic has to be familiar to the students, so that a lack of background knowledge or variability in background knowledge across students does not impact the validity of score. Creating sufficient alternate forms with topics of equivalent difficulty is especially challenging for measure development.

Recent work has resulted in several sets of discourse-level measures that address these challenges. For example, the prompts have been evaluated to ensure that they support children writing about familiar topics, so that individual differences in background knowledge or life experiences do not negatively affect text generation. Children can select topics (and vocabulary) that are familiar, which could support transcription of their ideas into written text. However, open-ended prompts often yield responses that widely vary across children. In response to a prompt such as *One day, I was walking home from school and ...* an individual child could write a personal narrative about what happens each day when she walks home, a personal narrative about an interesting walk home one day last year, or a fantasy about meeting a pink dragon while walking home one day. Or a child such as Toby may write simple sentences such as "I like school" or "I saw my friend." The ability to write on any topic might support fluent writing or might hinder it, especially if children have difficulty planning or generating ideas. As an alternative, tasks that provide more specific topical focus such as a specific story might provide approaches to evaluate the content of what was written or support children who have difficulty with text generation.

A limitation of discourse-level measures as indicators of fluent writing is that children, especially young children, do not produce a great deal of text. If children

are exerting a great deal of cognitive resources into generating and transcribing ideas, text production may be limited given the allotted writing time. Efficient administration is a key characteristic of CBM, but children write slowly and produce little text in 2 or 3 min. This can impact the reliability, validity, and utility of scores to inform instructional practices for progress monitoring. Increasing the duration of writing time or including multiple items (e.g., three stories instead of one) might improve technical features of the measures, but could become both fatiguing for students and cumbersome for teachers to administer and score. Clearly, discourse-level CBM writing measures also require future research to provide clearer expectations for growth and benchmarking. This research agenda will be increasingly important in light of the Common Core State Standards and the need for measures to assess writing progress.

## **Expanding Beyond Production-Based Definitions of Fluent Writing**

It is important to consider that quantitative assessment of writing does not capture some important dimensions of writing and that given the complexity of writing, in the broader research on writing and writing development studies have evaluated students' writing in terms of quality and productivity. Studies have shown that writing productivity is moderately related to writing quality for children in elementary grades (Abbott & Berninger, 1993; Kim, Al Otaiba et al., 2014; Olinghouse, 2008). While productivity and quality have been used widely as separate constructs in previous studies (Abbott & Berninger, 1993; Olinghouse, 2008; Olinghouse & Graham, 2009), a recent study with first-grade students has confirmed that the writing quality and productivity are separate dimensions (Kim et al., 2014).

Using data from first-grade students who were assessed using a timed prompt, children's scores on the ideation, organization, sentence structure, and vocabulary choice aspects have been shown to capture a construct of writing quality. In contrast, the number of words, number of different words, and number of ideas were shown to capture the distinct construct of writing productivity. Below is a description of how writing quality and productivity have been conceptualized and operationalized.

### ***Writing Quality***

Writing quality is typically conceptualized as the extent to which ideas (or topics) are developed, and how those ideas are expressed and organized. Previous studies have measured writing quality in the following three ways. First, a holistic rating scale (e.g., 1–7 or 1–8) is used to evaluate various aspects of students' writing in such as idea (ideation), organization, grammar, and word choice with similar weight on all the aspects (Graham et al., 2002; Graham, Harris, & Mason, 2005; Graham,

Berninger, & Fan, 2007; Olinghouse, 2008). A rater rates the student's composition on an overall impression of quality taking into accounting these multiple aspects. Similarly, Babayigit and Stainthrop (2011) considered accuracy and clarity of the depiction of the events in the pictures and the appropriate use of vocabulary for writing quality.

The second way is rating the student's composition on a rating scale on predetermined aspects related to quality such as ideation, organization, sentence structure, and vocabulary. For narrative stories, inclusion of specific narrative elements is also rated (Hooper, Swartz, Wakelly, de Kruif, & Montgomery, 2002). This second approach is different from the first approach in that a rating score is available on each aspect rather than a single holistic score. For instance, in Olinghouse and Graham's (2009) study, students' narrative stories were evaluated on a 7-point scale in three aspects (or traits): (1) organization, (2) the development of plot, characters, and setting, and (3) creativity. Likewise, Wagner and colleagues (2011) examined students' written composition on an expository prompt in terms of presence of topic sentence, logical ordering of ideas, and presence of main idea, body, and conclusion.

Coker and Ritchey investigated qualitative scoring procedures for two measures: Sentence Writing and Picture Story. This score was developed to address limitations in scores that only evaluate production and to provide a score that reflects possible instructional goals and objectives, especially for younger students who would have difficulty producing text with sufficient length to analyze. For Sentence Writing Coker and Ritchey, (2010), the qualitative score is a composite of ratings of five unique components of writing: response type (if the student's response is a word or words, a complete sentence, or multiple sentences), spelling (percentage of words spelled correctly), mechanics (capitalization and punctuation), grammatical structure, and relation to prompt (the extent to which the student's response was related or unrelated to the prompt topic). The qualitative score for Sentence Writing had internal consistency of approximately .80 and alternate-form reliability of  $r = .65-.71$  (which is a slightly smaller magnitude in comparison to the production-based scores of total words written, correctly spelled words, and correct word sequences,  $r = .74-.81$ ). The concurrent criterion-related validity coefficients was approximately  $r = .40$  for kindergarten and  $r = .50$  for first-grade students.

The qualitative score for Picture Story was developed in a similar manner (Ritchey & Coker, 2013), but includes expanded components, as the writing task and the grade levels are different (first to third grade). In addition to scoring for response type, spelling, mechanics, grammatical structure, and relation to the prompt criteria included from Sentence Writing, components were added for descriptive words and the criteria for scores within other components (e.g., response type) were revised to reflect more advanced writing skills. The qualitative score for Picture Story had internal consistency that ranged from  $r = .77$  to  $.80$ , and concurrent criterion-related validity coefficients using WJ-III Writing Samples (Woodcock et al., 2001, 2007) as the criterion was  $r = .44$  for the sample of second- and third-grade students.

This recent work suggests that it may be possible to extend scoring procedures to include aspects of quality and to include attention to the content written by students



(Tindal & Hasbrouk, 1991). This approach to scoring writing samples also aligns to current practices, such as use of holistic and analytic writing rubrics, and makes them appealing to classroom teachers, particularly because qualitative information can provide diagnostic information that is useful for determining what to focus on during instruction. Further research is needed to determine whether such information does, indeed, enhance instruction and improve student outcomes. Further research is also needed to determine how genre may be important as a consideration in evaluating writing quality, especially given broad coverage of genres in the Common Core State Standards. It could be that other options are needed to address other aspects of writing that are important for writing across genres.

## Correlates of Writing that May Inform Instructional Targets

Just as understanding the correlates and predictive componential skills of reading has led to a better understanding of interventions, there is a need to refine such knowledge about predictors, correlates, and componential skills of writing. This knowledge could provide targets for identification of important areas for assessment and instruction in writing and lead to better understanding of how to meet the needs of students who experience difficulty in writing. Further, this knowledge will be needed to understand whether students are meeting Common Core State Standards for writing.

For example, Kim, Al Otaiba, Puranik, Sidler, Greulich, and Wagner (2011) used structural equation modeling (SEM) to investigate the shared and unique relations of potential component skills of writing of 242 kindergarten beginning writers (i.e., beginning composition). At the end of the school year, research assistants introduced the writing task and attempted to orient children to task expectations through a brief group discussion, as follows: *You have been in kindergarten for almost a whole year. Today we are going to write about kindergarten. Let's think about what you enjoyed about being in kindergarten. What did you learn in school? Did anything special happen to you in kindergarten?* Children had 15 min to write. Using the coding scheme developed by Puranik, Lombardino, and Altmann (2007; 2008), three variables were derived from students' writing: words written, number of ideas, and number of sentences. The authors categorized words as correct that were recognizable in the context of the child's writing despite some spelling errors. By contrast, random strings of letters or sequences of nonsense words (both were very rare in the sample) were not counted as words. Number of ideas was a count of the total number of propositions (i.e., predicate and argument) included in the child's writing sample. For example, *I love kindergarten* was counted as one idea. Finally, sentences was the count of the number of sentences included in the writing sample. Sentence structure was used to determine the number of sentences when punctuation and capitalization were not used, which is common for kindergartners. These scores were used to create a latent variable of writing performance.

The findings indicated that oral language, spelling, and letter writing automaticity were uniquely related to end of kindergarten writing performance. Moreover, variation in students' reading skills was not significantly related to their writing performance after these three component skills were entered into the model. This is likely because the contribution of reading to writing, at this early stage of development, may have been explained by spelling, as indicated by the strong correlation between the spelling and reading latent variables ( $r = .74$ ). Indeed, individual differences in spelling and letter writing fluency were uniquely related to beginning writing, but these components were related to different aspects of writing. Spelling appeared to capture phonological, alphabetic, and orthographic knowledge (Cassar, Treiman, Moats, Pollo, & Kessler, 2005; Kim, 2010; Moats, 2005/2006), whereas letter writing fluency may have captured students' knowledge of letters and the ability to write them efficiently. The findings also emphasize the importance of oral language skills (composed of vocabulary, grammatical knowledge, and sentence imitation), which showed a moderate bivariate correlation with writing ( $r = .41$ ). Kim et al. (2011) suggested that students with stronger spelling and handwriting skills might be better able to use their oral language skills to devote attention and working memory to various higher-order aspects of writing (e.g., planning, translating, and revising).

Another recent study has examined a component that has been less commonly examined in writing fluency studies: self-regulation. Kent and colleagues (Kent, Wanzek, Petscher, Al Otaiba, & Kim, 2014) examined the unique and shared role of self-regulation, transcription, reading, and language ability longitudinally across kindergarten and first grade. The authors formed latent variables and used SEM to demonstrate that a model including self-regulation was better fitting than a model with only language and literacy factors. Three factors were uniquely and positively related to compositional fluency in kindergarten: self-regulation, reading and spelling proficiency, and letter writing automaticity. For first grade, the self-regulation and higher-order literacy factors were predictive of both composition quality and fluency and oral language showed unique relations only with first grade writing quality. Findings from these two studies suggest that it is important to attend not only to transcription skills but also to increase students' self-regulation.

In a similar study, Kim, Al Otaiba, Folsom, Greulich, and Puranik (2014) used a variety of scoring approaches (modified 6+1 trait scoring, Education Northwest, 2012), syntactic complexity measures, productivity measures such as number of words and ideas) to identify writing dimensions and then examined the shared and unique relations of oral language and literacy skills (i.e., reading, spelling, and letter writing fluency) to the identified dimensions of written composition. First-grade students ( $N = 527$ ) were assessed on oral language (vocabulary and grammatical knowledge), reading (word reading and oral reading fluency), spelling, letter writing fluency, and writing in the spring. Kim and colleagues used a latent variable approach and SEM. They found that the seven traits in the 6+1 trait system represented two constructs, which they called "substantive quality" and "spelling and writing conventions." In contrast, when the other scoring procedures were included (productivity and syntactic complexity), four dimensions emerged: substantive quality, productivity, syntactic complexity, and spelling and writing conventions.



They reported that language and literacy predictors were differentially related to each of these four dimensions of written composition.

In a study with children in second and third grade, unique predictors differed for different writing outcomes (Kim, Al Otaiba, Wanzek, & Gatlin, 2015). Specifically, when children's CBM scores (i.e., a latent variable of correct minus incorrect word sequences and percent of correct word sequences) in response to three prompts were the outcome, children's reading skill, spelling, paragraph copying, and attentiveness were unique predictors. In contrast, when the outcome was writing quality, oral language, letter writing automaticity, and rapid automatized naming were additional unique predictors. In other words, different dimensions or aspects of writing draw from different sets of language, literacy, and cognitive skills.

Finally, gender appears to play a role in writing achievement. Ever since writing has been included in the National Assessment of Educational Progress (NAEP), girls have outperformed boys consistently across grades 4, 8, and 12 with effect sizes greater than 0.5 (e.g., National Center for Education Statistics, 2003). This gap in writing is even found with beginning writers (Kim et al., 2013; Knudson, 1995), particularly with respect to performance level (but not necessarily growth rate; Parker, McMaster, Medhanie, & Silberglitt, 2011). Despite this consistent gap in writing as a function of gender, our understanding of sources of gaps is limited. A few studies suggest that attitude toward writing may be one explanation as boys tend to have less positive attitudes toward writing and less value in writing than girls (Knudson, 1995; Lee, 2013; but also see Graham, Berninger, & Fan, 2007). Additionally, differences in oral language, literacy, and attentiveness partially explained gender differences in writing (Kim et al., 2015).

In summary, important correlates and predictors for fluent writing in the early years appear to converge with theoretical models. Transcription skills and letter writing fluency appear to be important predictors, as is self-regulation. In addition, these variables predict both fluent writing and the quality of writing in terms of ideation, syntactic complexity, and spelling and writing conventions. An obvious implication is that teachers not only consider both fluent writing but also qualitative features in their writing assessment and instruction. We emphasize, however, that additional research is needed to more accurately guide teachers in planning instruction and intervention.

## **Informing Classroom Instruction and Interventions**

Thus far, we have focused on the critical components of writing and how they might be assessed, with an eye toward identifying global indicators in line with a CBM framework. The original purpose of CBM was to provide educators with timely information about students' proficiency level and progress in core academic skills, such that problems could be identified early on and meaningful solutions could be identified and implemented (Deno, 2005). Consistent with this initial purpose, a major goal in developing CBM for young writers is to design an instructionally useful tool for identifying and responding to struggling young writers' difficulties. In

this section, we consider the current context of writing instruction in primary classrooms. Then, we propose an approach to integrating assessment and instruction to supporting young children's development of fluent writing using CBM.

### ***Current Knowledge About Writing Instruction and Intervention in the Primary Grades***

Starting in the early years of schooling, children's classroom experiences should include explicit instruction in basic writing skills. Given that writing requires the simultaneous management and coordination of multiple cognitive–linguistic processes (e.g., Berninger, 2008; Moats, 2005/2006), such instruction should emphasize each of the component processes of writing (as illustrated above). For example, instruction targeting automaticity of handwriting and spelling has led to improvements in young students' writing fluency and quality (e.g., Berninger et al., 2002, 2006; Graham, Harris, & Fink, 2000; Graham et al., 2002; Graham, McKown, Kiu-hara, & Harris, 2012; Jones & Christensen, 1999). At the same time, early instruction in generating text (including instruction in language use, text structure, genre, and content; Coker, 2006; Graham, McKeown, Kiu-hara, & Harris, 2011; Olinghouse & Leird, 2009), as well as self-regulation of the writing process (e.g., planning and organizing, reviewing and revising; e.g., Baker, Chard, Ketterlin-Geller, Apichatabutra, & Doabler, 2009; Harris, Graham, & Mason, 2006), is critical for developing overall writing proficiency.

Researchers have recommended that as beginning writers benefit more from short but frequent practice (Graham, 2006); Graham & Miller, 1980, writing instruction should be implemented daily, for about 30 min, starting in kindergarten (Edwards, 2003; Graham, Bollinger et al., 2012; Jones & Christensen, 1999). Further, instruction should include a balance between teacher instruction and student independent writing (e.g., Cutler & Graham, 2008; Gilbert & Graham, 2010). The Common Core State Standards further delineate what is expected of children as they complete each grade level (see Table 2.1).

Despite these recommendations, evidence suggests that teachers spend little time providing explicit early writing instruction. In survey studies, teachers have reported spending about 20 min a day on writing in the primary grades, with 90% reporting teaching an average of 70 min of handwriting instruction per week (Cutler & Graham, 2008; Graham et al., 2008). Though teachers report using a combination of instruction in the writing process and direct skills related to spelling, grammar, capitalization, and punctuation (Cutler & Graham, 2008), observational studies reveal a somewhat different picture with respect to the amount of time spent in meaningful writing instruction.

For example, Puranik, Al Otaiba, Folsom, and Greulich (2014) examined the nature of writing instruction in kindergarten classrooms and described student writing outcomes at the end of the school year. Twenty-one teachers and 238 kindergarteners from 9 schools participated. The entire 90 min instructional block for language arts was videotaped twice per year, and at the end of the year students completed handwriting fluency, spelling, and writing tasks. Findings indicated that, on average, teachers

spent only 6.1 min in the fall and 10.5 min in the winter on any kind of writing instruction. Students spent most of the time writing independently (journals, worksheets) and received very little modeling or scaffolding by their teachers. Approximately 2 min of spelling instruction, on average, was observed and only a minimal amount of time was spent on instruction focused on the writing process. Large variability was observed in the amount of writing instruction, the amount of time kindergarten teachers spent on writing, and the amount of time students spent on writing.

Marked variability in classroom practices both within and across schools was reflected in the large variability noted in kindergartners' handwriting, spelling, and writing performance (Puranik et al., 2014). For handwriting fluency, students were asked to write their *ABC*'s for 1 min. On average, students could write 9.9 letters ( $SD=6.08$ , Range 0–26); however, 7 of the 238 students did not write a single letter and about 40 students wrote fewer than 5 letters. Performance differences were noted among classrooms with mean class scores on the handwriting fluency task ranging from a low of 3.75 letters to a high of 15.45 letters. Similarly, the mean spelling score was 49.01 ( $SD=20.38$ , Range 0–82 out of a possible 84). Several children only used initial and final letters to spell words, and about 5–20% of children either did not respond or wrote a random string of letters to spell words. However, the mean class scores varied from 17.80 to 62.47 points. The third type of data collected was a 15 min sample from writing prompt that asked students to tell what they had learned in kindergarten. The sample mean for words written was 14.37 ( $SD=15.62$ , Range=0–90). As with the handwriting fluency and spelling assessments, large variability was noted, with some students writing only a few words to one child writing 90 words. Again there was considerable variation at the classroom level with the words written class mean ranging from 1 to 51.38. An observational study suggests that students at risk for reading difficulties may have even fewer opportunities for writing instruction. Kent, Wanzek, and Al Otaiba (2012) observed kindergarten language arts instruction to explore literacy activities broadly, and more specifically, the amount and type of engagement in reading print for over 100 kindergarten students at-risk for reading difficulties during their general education classroom reading instruction. Minimal amounts of time were noted for vocabulary, fluency, writing and spelling.

### ***Integrating Assessment and Instruction to Improve Fluent Writing in Young Children***

Converging evidence from studies such as those described above suggests that insufficient time and emphasis is given to writing instruction in classrooms. We propose that the current state of early writing assessment and instruction can and should be improved, and that one way of doing so is to integrate the types of assessments discussed in this chapter with instruction, such that teachers have a framework within which to make data-based decisions about the effectiveness of their instruction. Below, we describe how CBM can be used to inform instructional decisions. Then, we return to our case example to illustrate how data-based instructional decision-making could be part of a multitiered system of instructional supports to improve writing outcomes for young students.

## Using CBM to Inform Instructional Decisions

The fluency-based assessments described in this chapter are particularly useful for informing instruction because they directly measure students' skills within a curricular domain (Shapiro, 2011). Direct assessments use standardized procedures that produce reliable and valid data, such as those that have been described in this chapter. For example, students who read very few words correct per minute using direct assessment procedures for oral reading fluency are identified as needing additional support. Students who can complete only single-digit addition math facts in third grade also need support. Similarly, students who cannot correctly produce accurate letter formations and basic spelling, such as those described in the Puranik and Al Otaiba (2012) study, likely need additional support in writing. What follows are two examples of how CBM assessments have been used to target and evaluate early writing interventions.

The Picture-Word assessment was used in a recent application of brief experimental analysis (BEA) for beginning writing. BEA is an approach to direct assessment that tests the effects of different instructional variables on academic performance (Jones & Wickstrom, 2002). The instructional variables that are tested using BEA derive from hypotheses for why students struggle to perform academic skills, such as not wanting to complete the task (i.e., motivation), lack of practice, or not having had enough instruction (Daly, Witt, Martens, & Dool, 1997). What is necessary for BEA are direct assessments that provide information regarding the effects of instructional approaches designed to test these hypotheses. In reading, CBM was used to compare the effects of multiple interventions, and results showed clear differentiation in reading performance, which identified the most promising instructional approach (Daly, Martens, Dool, & Hintze, 1998; Hintze, 1998).

Parker, Dickey, Burns, and McMaster (2012) extended BEA procedures to struggling first-grade writers, and used CBM assessments to determine the effects of the tested hypotheses. Results of the BEA showed differentiation across the tested hypotheses for each participating student. As a follow-up, the same CBM procedures were used to assess the effects of the interventions that were implemented following the BEA. Each student's level and trend improved from baseline to intervention.

The Picture-Word assessment also shows promise for use in a skill-by-treatment interaction framework for targeting intervention (Burns, Coddling, Boice, & Lukito, 2010). Central to this framework is the idea that instructional approaches should be tailored to address students' present skill level in a way informed by concepts such as the zone of proximal development (Vygotsky, 1978) and the instructional level (Betts, 1945). The skill-by-treatment interaction framework operationalizes these concepts using a research-based model of skill development, in which skills develop according to stages of acquisition, fluency, generalization, and adaptation (Haring & Eaton, 1978). Depending on students' skill performance, different interventions may be necessary for different students.

Fluency-based direct assessments are promising for use within a skill-by-treatment interaction framework because they provide data for making inferences about whether fluency is (a) not an appropriate target because skills that are still being acquired, (b) an appropriate target because students are slow but accurate, or (c) sufficiently developed such that more complex skills can be targeted. Parker, Mc-

Master, and Burns (2011) used an extant data set to derive an instructional level for early writing skills performed by first-grade students, and the results showed promising technical characteristics for instructional-level criteria. Thus, students performing below the instructional level could be provided intervention that focuses on the accurate production of early writing skills, students performing within the instructional level could be provided intervention that gives additional practice and feedback, and students performing above the instructional level could be provided intervention that targets more complex, self-regulatory strategies (e.g., Graham & Harris, 1996). Applications of the skill-by-treatment interaction framework produced positive outcomes in reading (Parker & Burns, 2014) and in a meta-analysis of math interventions (Burns et al., 2010).

### **Using CBM in a Multitiered System of Support to Promote Fluent Writing in Young Children**

As illustrated above, CBM can be used to inform early writing instruction, although clearly further research is needed. Such data-based decision-making approaches are consistent with current multitiered systems of support that are used to ensure that all students are provided with appropriate instruction based on their needs (Gersten et al., 2008). One such multitiered framework is RTI, which includes (a) Tier 1: universal screening and evidence-based core instruction implemented with fidelity, (b) Tier 2: ongoing progress monitoring and supplementary instruction for students identified as at risk during screening, and (c) Tier 3: more intensive, individualized intervention for students for whom supplementary intervention is not sufficient and a corresponding increase in the specificity and frequency of progress assessment. RTI has been applied most broadly in reading (Gersten et al., 2008), and to a lesser extent, in math (Gersten et al., 2009; Lembke, Hampton, & Beyers, 2012). The measures discussed in this chapter provide preliminary direction for establishing an RTI approach to improving students' early writing outcomes. Continued research is needed to establish which of these assessments are best suited for screening and monitoring progress of early writing; thus, this proposal should be viewed as heuristic in nature (see also McMaster, Parker, & Jung, 2012; Saddler & Asaro-Saddler, 2013 for additional discussions of RTI applications to writing). Below, we illustrate how an RTI model could be used to address Toby's lack of fluent writing skills.

#### **Tier 1: Universal Screening and Core Instruction**

Recall that Toby's teacher, Mrs. Wright, has some concern about Toby's writing performance in her first-grade classroom. As part of her school's RTI process, she administers a CBM prompt to all of her students in fall of the school year, records each child's score, and notes the mean score of the class. Six students, including Toby, score 0 (far below the class mean of 6), so Mrs. Wright decides to monitor their progress on a weekly basis. Figure 2.3 shows the class average and the six at-risk students' baseline scores. The last data point on the graph shows the expected

first-grade CBM performance in winter, based on school-wide data. Mrs. Wright then begins her core language arts instruction, which includes 30 min of explicit daily handwriting, spelling, and composition instruction conducted with the whole class. As Fig. 2.3 shows, two students appear to make progress toward the expected winter benchmark, but four students do not (including Toby).

Tier 2: Supplementary Intervention

Mrs. Wright consults her school’s data team, which consists of the other first-grade teacher, the intervention specialist, a special education teacher, and a school psychologist. Based on their examination of the four students’ CBM samples, along with other products such as handwriting, spelling, and journal-writing samples, the team hypothesizes that though the students have no trouble generating ideas for writing, they have difficulties with transcription. They determine that an important first step will be to improve the students’ handwriting and spelling skills so that they can better transcribe their ideas onto paper. So, in addition to core instruction, the intervention specialist meets with the four students in a small group for 20 min daily, and implements explicit, evidence-based handwriting and spelling instruction. Three of the students respond well to the Tier 2 supplemental intervention, as evidenced by clear changes in their CBM slopes (see Fig. 2.3). However, Toby does not appear to be making progress. After 6 weeks of Tier 2 intervention, the data team determines that more intensive, individualized instruction is needed.

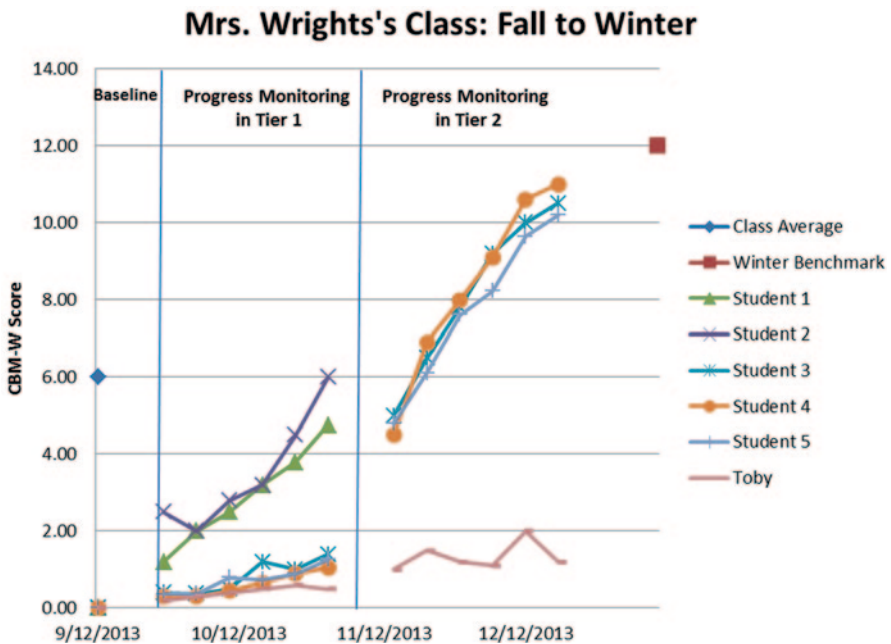


Fig. 2.3 Class data



### Tier 3: Highly Intensive, Individualized Intervention

In January, the intervention specialist, Mr. Lein, begins working with Toby for 30 min per day, 5 days per week. To identify an intervention to help Toby improve his writing skills, Mr. Lein conducts a series of brief, systematic tests of hypotheses for why Toby is struggling, using the BEA approach described above (Daly et al., 1997; Parker et al., 2012). Mr. Lein determines that a combination of modeling and repeated practice results in Toby’s best performance. Therefore, Mr. Lein adds modeling and repeated practice to the handwriting and spelling instruction already in place.

The data team sets a goal for Toby to improve his writing by one word spelled correctly per week on Picture-Word task, or 22 words spelled correctly by the end of the year, as indicated by the goal line on the graph in Fig. 2.4. As can be seen in Fig. 2.4, Toby begins to make progress, but not enough to meet the goal. Although his spelling and handwriting begin to improve, he continues to write very slowly. He does begin to use more descriptive words, such as “raptor” and “sedimentary.” When four of Toby’s data points fall below the goal line, Mr. Lein again implements the BEA procedure to decide on an appropriate instructional change. This time, Mr. Lein determines that adding a timed fluency-building component to Toby’s practice sessions will likely improve his progress. In fact, Toby does make more progress, but continues to fall short of his goal. Once again, the Mr. Lein uses BEA to identify another instructional change: daily goal-setting during the fluency-building activity with incentives for meeting daily goals. After this addition, Toby makes steady

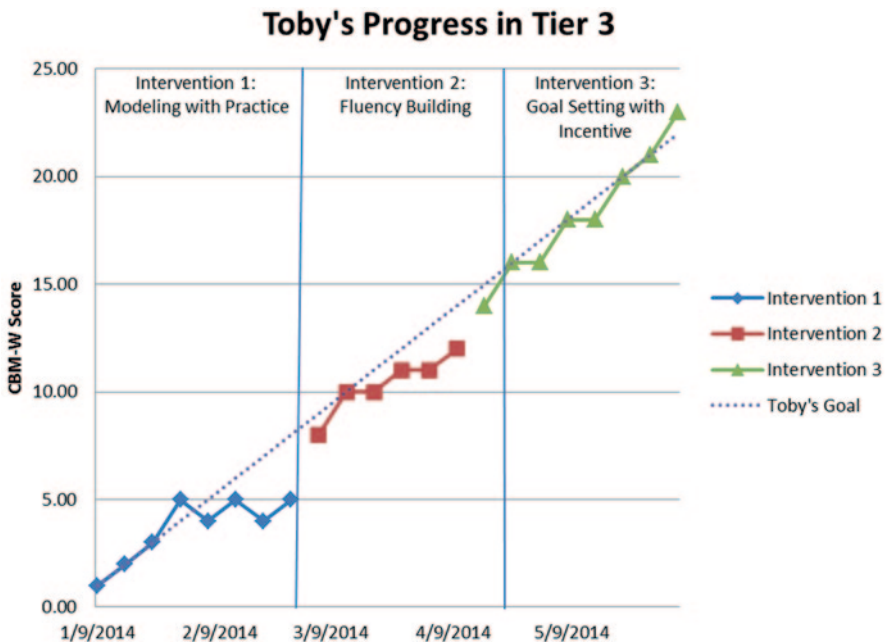


Fig. 2.4 Toby’s progress in Tier 3

progress, and by the end of the year, meets (and even exceeds) the goal, producing longer, more legible, accurate, and interesting texts.

If Toby continues to make good progress, he could receive instruction at Tier 1 and/or 2; if needed, he would again have access to Tier 3. The decision about the appropriate tier of instruction should be made based on ongoing progress-monitoring data. In this way, CBM remains an essential component in ensuring that Toby receives instruction that best meets his needs within an RTI framework.

## Directions for Future Research

Although, much work has been conducted focusing on the development and validation of indicators of fluent writing and on efforts to improve writing instruction, more work is needed. Much of the existing work has focused on identifying a parsimonious set of tasks that would be appropriate to capture fluent writing development for a specific writing area or grade-level group. In recent reviews (McMaster & Espin, 2007; McMaster, Ritchey, & Lembke, 2011), the technical features have been described with the goal of selecting those assessments that exhibit the strongest technical adequacy. More recent work has also tried to link the assessments to emerging knowledge and theoretical models of writing development. However, more research is needed to provide researchers and teachers with a broad set of assessments that can be used within and across grade levels. To date, there are still gaps in measures that align to developmental progressions and have evidence of technical adequacy within and across grades. For example, more work has been conducted in the spring of kindergarten and first grade, with a gap for appropriate fall measures. The majority of existing studies have been at a single grade or cross sectional; longitudinal investigation of children's performance on these measures over time holds promise for better understanding the components the a seamless assessment system.

We also propose that additional research is needed to evaluate the best way to capture students' writing using validated scoring techniques. Production scores have worked well as indicators of performance, but serve as a limited assessment of a child's overall writing proficiency. Future attention to the content and the quality of the child's writing could improve the utility of measures. In addition to the scoring construct, there are many directions for future research related to scoring efficiency. Scoring a class of children's writing samples each week using a variety of scores is time consuming, and teachers may lack the resources needed to quickly and reliably score large amounts of writing assessments. Technology-based scoring is one direction for future research.

Technology also holds promise for better understanding of children's fluent writing while engaged in writing activities. Touch screen technology or other ways to capture the online processes a child uses while writing could refine measurement procedures and support understanding of how children approach the writing stimuli included in these assessments. Technology may also provide data about individual's response time—especially about time to write letters, spell words, “thinking time”



involved in text generation, and other important online activities related to the role of self-regulation in fluent writing.

Finally, more work is needed with respect to the capacity of the measures to be useful for screening and progress monitoring. One of the goals of CBM is systematic universal screening that can occur several times per year. Ritchey and Coker have conducted short-term screening and classification accuracy studies (predicting from mid-year to the end of the year) using measures for students in kindergarten (Coker & Ritchey, 2014), first grade (Coker & Ritchey, 2012; Ritchey & Coker 2014), and second and third grades (Ritchey & Coker, 2013). In a similar way, additional research could extend the understanding of fluency-based assessments within the skill-by-treatment interaction framework (Burns et al., 2010) beyond the current state of understanding that exists only for first grade (Parker, McMaster & Burns, 2011). With respect to progress monitoring, most research has focused on whether the scores are sensitive to growth in short time periods, with small change in scores (Coker & Ritchey, 2010; McMaster, Ritchey, & Lemke, 2011). If these measures are to hold promise for monitoring progress, especially for students who are receiving supplemental instruction (i.e., Tier 2 and Tier 3), additional work is needed to develop measures which yield scores that are sensitive to weekly or biweekly growth. Last, it is important to learn more about how using these measures could lead to improved education outcomes.

## Conclusion

Learning to write is an important component of academic development. As illustrated in the chapter, multiple aspects of writing instruction can support development of fluent writing, including attention to the amount and quality of writing instruction and the use of assessments to guide instructional decisions. For children like Toby, these components are essential if his writing proficiency is to improve. Progress has been made in the development of technically adequate assessments and there is better understanding of the correlates of writing development and role of instruction in supporting this development. However, much more work is needed to develop a seamless system of instruction and assessments that can support the development of fluent writing for children in prekindergarten to third grade.

## References

- Abbott, R. D., & Berninger, V. W. (1993). Structural equation modeling of relationships among development skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology, 85*, 478–508.
- Babayigit, S., & Stainthorp, R. (2011). Modeling the relationships between cognitive-linguistic skills and literacy skills: New insights from a transparent orthography. *Journal of Educational Psychology, 103*, 169–189.
- Baker, S. K., Chard, D. J., Ketterlin-Geller, L., Apichatabutra, C., & Doabler, C. (2009). Teaching writing to at-risk students: The quality of evidence for self-regulated strategy development. *Exceptional Children, 75*(3), 303–318.

- Berninger, V. W. (1999). Coordinating transcription and text generation in working memory during composing: Automatized and constructive processes. *Learning Disability Quarterly*, 22, 99–112.
- Berninger, V. W. (2008). Defining and differentiating dysgraphia, dyslexia, and language learning disability within a working memory model. In M. Mody & E. R. Silliman (Eds.), *Brain, behavior, and learning in language and reading disorders* (pp. 103–134). New York: Guilford Press.
- Berninger, V. W. (2009). Highlights of programmatic, interdisciplinary research on writing. *Learning Disabilities Research & Practice*, 24, 69–80.
- Berninger, V. W., & Rutberg, J. (1992). Relationship of finger function to beginning writing: Application to diagnosis of writing disabilities. *Developmental Medicine and Child Neurology*, 34(3), 198–215.
- Berninger, V. W., & Swanson, H. L. (1994). Modifying Hayes and Flower's model of skilled writing to explain beginning and developing writers. In E. Butterfield (Ed.), *Children's writing: Toward a process theory of development of skilled writing* (pp. 57–81). Greenwich: JAI Press.
- Berninger, V., Abbott, R., Rogan, L., Reed, E., Abbott, S., Brooks, A., & And, O. (1998). Teaching spelling to children with specific learning disabilities: The mind's ear and eye beat the computer or pencil. *Learning Disability Quarterly*, 21(2), 106–122.
- Berninger, V. W., Vaughan, K., Abbott, R. D., Begay, K., Coleman, K., Curtain, G., & Graham, S. (2002). Teaching spelling and composition alone and together: Implications for the simple view of writing. *Journal of Educational Psychology*, 94, 291–304.
- Berninger, V., & Amtmann, D. (2003). Preventing written expression disabilities through early and continuing assessment and intervention for handwriting and/or spelling problems: Research into practice. In H. L. Swanson, K. Harris, & S. Graham (Eds.), *Handbook of research on learning disabilities* (pp. 345–363). New York: Guilford.
- Berninger, V. W., Rutberg, J. E., Abbott, R. D., Garcia, N., Anderson-Youngstrom, M., Brooks, A., et al. (2006). Tier 1 and tier 2 early intervention for handwriting and composing. *Journal of School Psychology*, 44, 3–30.
- Berninger, V. W., Nielsen, K. H., Abbott, R. D., Wijsman, E., & Raskind, W. (2008). Writing problems in developmental dyslexia: Under-recognized and under-treated. *Journal of School Psychology*, 46, 1–21.
- Betts, E. A. (1945). *Foundations of reading instruction*. New York: American Book.
- Biancarosa, G., & Snow, C. E. (2004). *Reading Next—A vision for action and research in middle and high school literacy: A report to Carnegie Corporation of New York*. Alliance for Excellent Education.
- Bloodgood, J. (1999). What's in a name? Children's name writing and literacy acquisition. *Reading Research Quarterly*, 34, 342–367.
- Both-de Vries, A., & Bus, A. G. (2008). Name writing: A first step to phonetic writing? Does the name have a special role in understanding the symbolic function of writing? *Literacy Teaching and Learning*, 12, 37–55.
- Both-de Vries, A., & Bus, A. G. (2010). The proper name as starting point for basic reading skills. *Reading and Writing*, 23, 173–187.
- Brigance, A. (1999). *Comprehensive inventory of basic skills-revised*. North Billerica: Curriculum Associates.
- Burns, M. K., Coddling, R. S., Boice, C. H., & Lukito, G. (2010). Meta-analysis of acquisition and fluency math interventions with instructional and frustration level skills: Evidence for a skill-by-treatment interaction. *School Psychology Review*, 39, 69–83.
- Cassar, M., Treiman, R., Moats, L., Pollo, T., & Kessler, B. (2005). How do the spellings of children with dyslexia compare with those of nondyslexic children? *Reading and Writing: An Interdisciplinary Journal*, 18, 27–49.
- Coker, D. L. (2006). Impact of first-grade factors on the growth and outcomes of urban schoolchildren's primary-grade writing. *Journal of Educational Psychology*, 98, 471–488.
- Coker, D. L., & Ritchey, K. D. (2010). Curriculum based measurement of writing in kindergarten and first grade: An investigation of production and qualitative scores. *Exceptional Children*, 76, 175–193.

- Coker, D. L., & Ritchey, K. D. (2012, February). *Predicting writing difficulties in first grade: An investigation of writing and reading measures*. Presented at the Pacific Coast Research Conference, Coronado, California.
- Coker, D. L., & Ritchey, K. D. (2014). Universal screening for writing risk in kindergarten. *Assessment for Effective Intervention*, 39, 245–256. doi:10.1177/1534508413502389.
- Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology*, 100, 907–919.
- Daly, E. J., III, Witt, J. C., Martens, B. K., & Dool, E. J. (1997). A model for conducting a functional analysis of academic performance problems. *School Psychology Review*, 26, 554–574.
- Daly, E. J., III, Martens, B. K., Dool, E. J., & Hintze, J. M. (1998). Using brief functional analysis to select interventions for oral reading. *Journal of Behavioral Education*, 8, 203–218.
- Dellerman, P., Coirer, P., & Marchand, E. (1996). Planning and expertise in argumentative composition. In G. Rijlaarsdam, H. van den Bergh, & M. Couzijn (Eds.), *Theories, models, and methodology in writing research* (pp. 182–195). Amsterdam: Amsterdam University Press.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52, 219–232.
- Deno, S. L. (2005). Problem-solving assessment. In R. Brown-Chidsey (Ed.), *Assessment for intervention: A problem-solving approach* (pp. 10–40). New York: Guilford Press.
- Deno, L., & Marston, D. (2006). Curriculum-based measurement of oral reading: An indicator of growth in fluency. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about fluency instruction*. International Reading Association.
- Deno, S. L., Mirkin, P. I. C., Lowry, L., & Kuehnle, K. (1980). *Relationships among simple measures of spelling and performance on standardized achievement tests*. Research Report No. 21. University of Minnesota Institute for Research on Learning Disabilities.
- Deno, S. L., Mirkin, P., & Marston, D. (1980). *Relationships among simple measures of written expression and performance on standardized achievement tests* (Vol. IRLD-RR-22). University of Minnesota, Institute for Research on Learning Disabilities.
- Deno, S. L., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. (1982). *The use of standard tasks to measure achievement in reading, spelling and written expression: A normative and developmental study* (Vol. IRLD-RR-87). University of Minnesota, Institute for Research on Learning Disabilities.
- Deno, S. L., Mirkin, P., & Marston, D. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children Special Education and Pediatrics: A New Relationship*, 48, 368–371.
- Diamond, K., Gerde, H., & Powell, D. (2008). Development in early literacy skills during the pre-kindergarten year in Head Start: Relations between growth in children's writing and understanding of letters. *Early Childhood Research Quarterly*, 23, 467–478.
- Drouin, M., & Harmon, J. (2009). Name writing and letter knowledge in preschoolers: Incongruities in skills and the usefulness of name writing as a developmental indicator. *Early Childhood Research Quarterly*, 24, 263–270.
- Education Northwest. (2012). *6+1 Trait® Writing*. <http://educationnorthwest.org/resource/949>. Accessed 13 June 2012.
- Edwards, L. (2003). Writing instruction in kindergarten: Examining an emerging area of research for children with writing and reading difficulties. *Journal of Learning Disabilities*, 36, 136–148.
- Ehri, C. (1986). Sources of difficulty in learning to spell and read. In M. L. Wolraich & D. Routh (Eds.), *Advances in developmental and behavioral pediatrics*, Vol. 7 (pp. 121–195). Greenwich: JAI Press.
- Ehri, L. C., & McCormick, S. (1998). Phases of word learning: Implications for instruction with delayed and disabled readers. *Reading and Writing Quarterly: Overcoming Learning Difficulties*, 14(2), 135–164.
- Espin, C. A., Scierka, B. J., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading and Writing Quarterly: Overcoming Learning Difficulties*, 15, 5–27.

- Ferreiro, E., & Teberosky, A. (Eds.). (1982). *Literacy before schooling*. Exeter: Heinemann.
- Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *Journal of Special Education, 41*, 121–139.
- Fry, E. B., Kress, J. E., & Fountoukidis, D. L. (1993). *The reading teacher's book of lists*. Englewood Cliffs: Prentice Hall.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Allinder, R. M. (1991a). The contribution of skills analysis to curriculum-based measurement in spelling. *Exceptional Children, 57*, 443–452.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Allinder, R. M. (1991b). Effects of expert system advice within curriculum-based measurement in teacher planning and student achievement in spelling. *School Psychology Review, 20*, 49–66.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*, 27–48.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239–256.
- Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Slider, N. J., Hoffpauir, L. D., & Whitmarsh, E. L., et al. (2004). An examination of the criterion validity and sensitivity to brief intervention of alternate curriculum-based measures of writing skill. *Psychology in the Schools, 41*, 291–300.
- Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to intervention for elementary and middle schools*. (NCEE 2009-4060). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. <http://ies.ed.gov/ncee/wwc/publications/practiceguides/>. Accessed 1 June 2012.
- Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W. D. (2008). *Assisting students struggling with reading: Response to Intervention and multi-tier intervention for reading in the primary grades. A practice guide*. (NCEE 2009-4045). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. <http://ies.ed.gov/ncee/wwc/publications/practiceguides/>. Accessed 1 June 2012.
- Gilbert, J., & Graham, S. (2010). Teaching writing to elementary students in grades 4 to 6: A national survey. *Elementary School Journal, 110*, 494–518.
- Good, R. H., III, & Kaminski, R. A. (1996). Assessment for instructional decisions: Toward a proactive/ prevention model of decision-making for early literacy skills. *School Psychology Quarterly, 11*, 326–336.
- Graham, S. (2006). Strategy instruction and the teaching of writing. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187–207). New York: Guilford.
- Graham, S., & Harris, K. R. (1996). Addressing problems in attention, memory, and executive functioning: An example from self-regulated strategy development. In G. Lyon & N. A. Krasnegor (Eds.), *Attention, memory, and executive function* (pp. 349–365). Baltimore: Paul H Brookes Publishing.
- Graham, S., & Harris, K. R. (2005). Improving the writing performance of young struggling writers: Theoretical and programmatic research from the center on accelerating student learning. *Journal of Special Education, 39*, 19–33.
- Graham, S., & Hebert, M. (2010). *Writing to read: Evidence for how writing can improve reading (A report from the Carnegie Corporation of New York)*. Washington, DC: Alliance for Excellent Education.
- Graham, S., & Miller, L. (1980). Handwriting research and practice: A unified approach. *Focus on Exceptional Children, 13*, 1–16.
- Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high school*. Washington, DC: Alliance for Excellent Education.
- Graham, S., Berninger, V. W., Abbott, R., Abbott, S., & Whitaker, D. (1997). The role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology, 89*, 170–182.

- Graham, S., Harris, K. R., & Fink, B. (2002). Is handwriting causally related to learning to write? Treatment of handwriting problems in beginning writers. *Journal of Educational Psychology, 92*, 620–633.
- Graham, S., Harris, K. R., & Fink, B. F. (2000). Contribution of spelling instruction to the spelling, writing, and reading of poor spellers. *Journal of Educational Psychology, 94*, 669–686.
- Graham, S., Harris, K. R., Fink-Chorzempa, B., & MacArthur, C. (2003). Primary grade teachers' instructional adaptations for struggling writers: A national survey. *Journal of Educational Psychology, 95*, 279–292. doi:10.1037/0022-0663.95.2.279.
- Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, self-efficacy of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology, 30*, 207–241.
- Graham, S., Berninger, V. W., & Fan, W. (2007). The structural relationship between writing attitude and writing achievement in first and third grade students. *Contemporary Educational Psychology, 32*, 516–536.
- Graham, S., Bollinger, A., Olson, C., D'Aoust, C., MacArthur, C., McCutchen, D., & National Center for Education Evaluation and Regional Assistance (Ed.). (2012). *Teaching Elementary School Students to Be Effective Writers: A Practice Guide. NCEE 2012-4058*. What Works Clearinghouse.
- Graham, S., Harris, K. R., Mason, L., Fink-Chorzempa, B., Moran, S., & Saddler, B. (2008). How do primary grade teachers teach handwriting? A national survey. *Reading and Writing: An Interdisciplinary Journal, 21*, 49–69.
- Graham, S., McKeown, D., Kiuahara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*. doi:10.1037/a0029185.
- Hammill, D. D., & Larsen, S. C. (1996). *Test of Written Language-3*. Austin: Pro-Ed.
- Hampton, D., Lembke, E. S., & Summers, J. (2013). *Examining the technical adequacy of early writing curriculum-based progress monitoring measures*. Unpublished manuscript.
- Haney, M. (2002). Name writing: A window into the emergent literacy skills of young children. *Early Childhood Education Journal, 30*, 101–105.
- Haney, M., Bisonnette, V., & Behnken, K. L. (2003). The relationship between name writing and early literacy skills in kindergarten children. *Child Study Journal, 33*, 99–115.
- Haring, N. G., & Eaton, M. D. (1978). Systematic instructional technology: An instructional hierarchy. In N. G. Haring, T. C. Lovitt, M. D. Eaton, & C. L. Hansen (Eds.), *The fourth R: Research in the classroom*. Columbus: Merrill.
- Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Hillsdale: Erlbaum.
- Harris, A. P., & Jacobson, M. D. (1972). *Basic elementary reading vocabularies*. New York: Macmillan.
- Harris, K. R., Graham, S., & Mason, L. H. (2006). Improving the writing, knowledge, and motivation of struggling young writers: Effects of self-regulated strategy development with and without peer support. *American Educational Research Journal, 43*, 295–337.
- Hintze, J. M. (1998). Using brief functional analysis to select interventions for oral reading. *Journal of Behavioral Education, 8*, 203–218.
- Hresko, W., Herron, S., & Peak, P. (1996). *Test of Early Written Language* (2nd ed.). Austin: PRO-ED.
- Hresko, W., Herron, S., Peak, P., & Hicks, D. (2012). *Test of Early Written Language* (3rd ed.). Austin: PRO-ED.
- Hooper, S. R., Swartz, C. W., Wakely, M. B., de Kruif, R. E. L., & Montgomery, J. W. (2002). Executive function in elementary school children with and without problems in written expression. *Journal of Learning Disabilities, 35*, 57–68.
- Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a Response to Intervention framework. *School Psychology Review, 36*, 582–600.
- Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review, 34*, 27–44.



- Jones, D., & Christensen, C. A. (1999). Relationship between automaticity in handwriting and students' ability to generate written text. *Journal of Educational Psychology, 91*, 44–49.
- Jones, K. M., & Wickstrom, K. F. (2002). Done in sixty seconds: Further analysis of the brief assessment model for academic problems. *School Psychology Review, 31*, 554–568.
- Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology, 78*, 243–255.
- Kent, S., Wanzek, J., & Al Otaiba, S. (2012). Print reading in general education kindergarten Classrooms: What does it look like for students at-risk for reading difficulties? *Learning Disabilities Research and Practice, 27*(2), 56–65.
- Kent, S., Wanzek, J., Petscher, Y., Al Otaiba, S., & Kim, Y.-S. (2014). Writing fluency and quality in kindergarten and first grade: The role of self-regulation, reading, transcription, and oral language. *Reading and Writing: An Interdisciplinary Journal, 27*, 1163–1188. doi:10.1007/s11145-013-9480-1.
- Kim, Y. (2010). Componential skills in early spelling development in Korean. *Scientific Studies of Reading, 14*, 137–158.
- Kim, Y.-S., Al Otaiba, S., Folsom, J. S., Greulich, L., & Puranik, C. (2014). Evaluating the dimensionality of first-grade written composition. *Journal of Speech, Language, and Hearing Sciences, 57*, 199–211.
- Kim, Y.-S., Al Otaiba, S., Puranik, C., Sidler, J. F., Gruelich, L., & Wagner, R. K. (2011). Componential skills of beginning writing: An exploratory study at the end of kindergarten. *Learning and Individual Differences, 21*, 517–525.
- Kim, Y.-S., Al Otaiba, S., Wanzek, J., & Gatlin, B. (2015). Towards an understanding of dimensions, predictors, and gender gap in written composition. *Journal of Educational Psychology, 107*, 79–95.
- Knudson, R. (1995). Writing experiences, attitudes, and achievement of first to sixth graders. *Journal of Educational Research, 89*, 90–97.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293–323.
- Lee, J. (2013). Can writing attitudes and learning behavior overcome gender difference in writing? Evidence from NAEP. *Written Communication, 30*, 164–193.
- Lembke, E., Deno, S., & Hall, K. (2003). Identifying an indicator of growth in early writing proficiency for elementary school students. *Assessment for Effective Intervention, 28*, 23–35.
- Lembke, E., Hampton, D., & Beyers, S. J. (2012). Response to intervention in mathematics: Critical elements. *Psychology in the Schools, 49*, 257–272. doi:10.1002/pits.
- Levin, I., & Bus, A. G. (2003). How is emergent writing based on drawing? Analyses of children's products and their sorting by children and mothers. *Developmental Psychology, 39*, 891–905.
- Levin, I., Vries, A. B., Aram, D., & Bus, A. (2005). Writing starts with own name writing: From scribbling to conventional spelling in Israeli and Dutch children. *Applied Psycholinguistics, 26*(3), 463–478.
- Lonigan, C., Schatschneider, C., & Westberg, L. (2008). *Results of the national early literacy panel research synthesis: Identification of children's skills and abilities linked to later outcomes in reading, writing, and spelling* (pp. 55–106). Washington, DC: National Early Literacy Panel.
- Marston, D., & Deno, S. (1981). *The reliability of simple, direct measures of written expression* (Vol. IRLD-RR-50). U.S.; Minnesota.
- McCutchen, D. (2006). Cognitive factors in the development of children's writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of Writing Research* (pp. 115–130). New York: Guilford.
- McMaster, K. L., & Campbell, H. (2008). New and existing curriculum-based writing measures: Technical features within and across grades. *School Psychology Review, 37*(4), 550–566.
- McMaster, K. L., & Espin, C. (2007). Curriculum-based measurement in writing: A literature review. *Journal of Special Education, 41*, 68–84.
- McMaster, K. L., Du, X., & Petrusdotir, A. (2009). Technical features of curriculum-based measures for beginning writers. *Journal of Learning Disabilities, 42*, 41–60.

- McMaster, K. L., Du, X., Yeo, S., Deno, S. L., Parker, D., & Ellis, T. (2011). Curriculum-based measures of beginning writing: Technical features of the slope. *Exceptional Children, 77*, 185–206.
- McMaster, K., Ritchey, K. D., & Lembke, E. (2011). Curriculum-based measurement of elementary students' writing: Recent developments and future directions. In T. E. Scruggs & M. A. Mastropieri (Eds.), *Assessment and intervention: Advances in learning and behavioral disabilities* (vol. 24) (pp. 111–148). Bingley: Emerald.
- McMaster, K. L., Parker, D., & Jung, P. G. (2012). Using curriculum-based measurement for beginning writers within a response to intervention framework. *Reading Psychology, 33*(1–2), 190–216.
- Moats, L. C. (2005/2006). How spelling supports reading: And why it is more regular and predictable than you may think. *American Educator, 29*(4), 12–22, 42–43.
- Molfese, V., Beswick, J., Jacobi-Vessels, J., Armstrong, N., Culver, B., White, J., & Molfese, D. (2011). Evidence of alphabetic knowledge in writing: Connections to letter and word identification skills in preschool and kindergarten. *Reading and Writing: An Interdisciplinary Journal, 24*, 133–150.
- National Center for Education Statistics. (2003). *The Condition of Education 2003*. (NCES 2003-067). U.S. Department of Education. Washington, DC: U.S. Government Printing Office. [http://nces.ed.gov/programs/coe/2003/pdf/19\\_2003.pdf](http://nces.ed.gov/programs/coe/2003/pdf/19_2003.pdf).
- National Center for Education Statistics. (2012). *The nation's report card: Writing 2011* (NCES 2012-470). <http://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf>.
- National Commission on Writing (2003). The neglected “R:” The need for a writing revolution. <http://www.writingcommission.org/>. Accessed 16 March 2007.
- National Commission on Writing (2004). *Writing: A ticket to work ... or a ticket out: A survey of business leaders*. College Entrance Examination Board.
- National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects K-5*. National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C. [http://www.corestandards.org/assets/CCSSI\\_ELA%20Standards.pdf](http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf). Accessed 27 Nov 2012.
- Olinghouse, N. G. (2008). Student- and instruction-level predictors of narrative writing in third-grade students. *Reading and Writing: An Interdisciplinary Journal, 21*, 3–26.
- Olinghouse, N. G., & Graham, S. (2009). The relationship between discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology, 101*, 37–50.
- Olinghouse, N. G., & Leaird, J. T. (2009). The relationship between measures of vocabulary and narrative writing quality in second- and fourth-grade students. *Reading and Writing: An Interdisciplinary Journal, 22*, 545–565.
- Parker, D. C., McMaster, K. L., & Burns, M. K. (2011). Determining an instructional level for beginning early writing skills. *School Psychology Review, 40*, 158–167.
- Parker, D. C., McMaster, K. L., Medhanie, A., & Silbergitt, B. (2011). Modeling early writing growth with curriculum-based measures. *School Psychology Quarterly, 26*(4), 290–304. doi:10.1037/a0026833.
- Parker, D. C., Dickey, B. N., Burns, M. K., & McMaster, K. L. (2012). An application of brief experimental analysis with early writing. *Journal of Behavioral Education, 21*, 329–349. doi:10.1007/s10864-012-9151-3.
- Parker, D. C., & Burns, M. K. (2014). Using the instructional level as a criterion to target reading interventions. *Reading and Writing Quarterly, 31*, 56–67.
- Parker, R. I., Tindal, G., & Hasbrouck, J. (1991). Countable indices of writing quality: Their suitability for screening-eligibility decisions. *Exceptionality, 2*, 1–17.
- Puranik, C., & Lonigan, C. (2011). From scribbles to scrabble: Preschool children's developing knowledge of written language. *Reading and Writing: An Interdisciplinary Journal, 24*, 567–589.

- Puranik, C., & Lonigan, C. (2012). Name writing proficiency, not length of name, is associated with preschool children's emergent literacy skills. *Early Childhood Research Quarterly, 27*, 284–294. doi:10.1016/j.ecresq.2011.09.003.
- Puranik, C., Schreiber, S., Estabrook, E., & O'Donnell, E. (2014). Comparison of name writing rubrics: Is there a gold standard? *Assessment for Effective Intervention, 40*, 16–23.
- Puranik, C., Lombardino, L., & Altmann, L. (2007). Writing through retellings: An exploratory study of language-impaired and dyslexic populations. *Reading & Writing, 20*, 251–272. doi:10.1007/s11145-006-9030-1.
- Puranik, C., Lombardino, L., & Altmann, L. (2008). Assessing the microstructure of written language using a retelling paradigm. *American Journal of Speech Language Pathology, 17*, 107–120.
- Puranik, C., Petscher, Y., & Lonigan, C. (2012). Dimensionality and reliability of letter writing in 3- to 5-year-old preschool children. *Learning and Individual Differences*. doi:10.1016/j.lindif.2012.06.011.
- Puranik, C., Al Otaiba, S., Folsom, J. S., & Gruelich, L. (2014). Exploring the amount and type of writing instruction during language arts instruction in kindergarten classrooms. *Reading and Writing: An Interdisciplinary Journal, 27*, 213–236. doi:10.1007/s11145-013-9441-8.
- Reid, D. K., Hresko, W. P., & Hammill, D. D. (2001). *Test of Early Reading Ability* (3rd ed.). Austin: Pro Ed.
- Ritchey, K. D. (2006). Learning to write: Progress monitoring tools for beginning and at-risk writers. *Teaching Exceptional Children, 39*(2), 22–26.
- Ritchey, K. D., & Coker, D. L. (2013). A comparison of the validity and utility of two curriculum based measurement writing tasks. *Reading and Writing Quarterly: Overcoming Learning Difficulties, 29*, 89–119. doi:10.1080/10573569.2013.741957.
- Ritchey, K. D., & Coker, D. L. (2014). Identifying writing difficulties in first grade: An investigation of writing and reading measures. *Learning Disabilities Research and Practice, 39*, 245–256.
- Saddler, B., & Asaro-Saddler, K. (2013). Response to Intervention in writing: A suggested framework for screening, intervention, and progress monitoring. *Reading and Writing Quarterly, 29*, 20–43. doi:10.1080/10573569.2013.741945.
- Shanahan, T. (2004). Overcoming the dominance of communication: Writing to think and to learn. In T. L. Jetton & D. A. Dole (Eds.), *Adolescent research and practice* (pp. 59–73). New York: Guilford.
- Shapiro, E. S. (2011). *Academic skills problems: Direct assessment and intervention* (4th ed.). New York: Guilford Press.
- Shinn, M. R., & Shinn, M. M. (2002). *Administration and scoring of spelling curriculum based measurement (S-CBM) for use in general outcome measurement*. Eden Prairie: Edformation.
- Sulzby, E., Barnhart, J., & Hieshima, J. (1989). *Forms of writing and rereading from writing: A preliminary report* (Technical report No. 20). <http://www.writingproject.org/Resources.techreports.html>. Accessed 1 June 2012.
- Tindal, G., & Hasbrouck, J. (1991). Analyzing student writing to develop instructional strategies. *Learning Disabilities Research and Practice, 6*, 237–245.
- Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research & Practice, 6*(4), 211–218.
- Tolchinsky, L. (2006). The emergence of writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 83–95). New York: Guilford Press.
- Treiman, R., & Bourassa, D. (2000a). Children's written and oral spelling. *Applied Psycholinguistics, 21*, 183–204.
- Treiman, R., & Bourassa, D. C. (2000b). The development of spelling skill. *Topics in Language Disorders, 20*(3), 1–18.
- Treiman, R., & Broderick, V. (1998). What's in a name: Children's knowledge about the letters in their own names. *Journal of Experimental Child Psychology, 70*, 97–116.
- Treiman, R., Kessler, B., & Bourassa, D. (2001). Children's own name influences their spelling. *Applied Psycholinguistics, 22*, 555–580.



- Troia, G. A. (2006). Writing instruction for students with learning disabilities. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 324–336). New York: Guilford Press.
- Vanderhyden, A., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review, 30*, 363–382.
- Videen, J., Deno, S. L., & Marston, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression* (Vol. IRLD-RR-84). University of Minnesota, Institute for Research on Learning Disabilities.
- Vygotsky, L. S. (1978). Interaction between learning and development (M. Lopez-Morillas, Trans.). In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society: The development of higher psychological processes* (pp. 79–91). Cambridge: Harvard University Press.
- Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Tschinkel, E., & Kantor, P. T. (2011). Modeling the development of written language. *Reading and Writing: An Interdisciplinary Journal, 24*, 203–220.
- Wayman, M., Wallace, T., Wiley, H., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41*, 85–120.
- Welsh, J., Sullivan, A., & Justice, L. (2003). That's my letter! What preschoolers' name writing representations tell us about emergent literacy knowledge. *Journal of Literacy Research, 35*, 757–776.
- Whitaker, D., Berninger, V., Johnston, J., & Swanson, L. (1994). Intraindividual differences in level of language in intermediate grade writers: Implications for the translating process. *Learning and Individual Differences, 6*, 107–130.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001, 2007). *Woodcock-Johnson III Tests of Achievement*. Itasca: Riverside.

## Chapter 3

# Mathematics Fluency—More than the Weekly Timed Test

Ben Clarke, Nancy Nelson and Lina Shanley

Pencils up, start...tick...tick...tick, stop, pencils down. If there is a shared cornerstone experience in education, the weekly timed test may be the winner. But why? Why across decades have teachers spoken those words and students furiously worked through sheets containing a range of problems from addition facts to division facts? This chapter attempts to provide answers to that fundamental question. We start with an exploration as to why fluency in mathematics is critical, examine interventions designed to increase fluency, and in the end provide an overview of the measures used to assess fluency and provide our thoughts to guide future work as the field gains a greater understanding of mathematics fluency.

As a nation, we compete in an international marketplace driven by technological innovation. Employment projections by the US Bureau of Labor Statistics indicate that the majority of the fastest growing occupations in the coming decade will require substantial preparation in mathematics or science (Lockard & Wolf, 2012). As policy-makers seek to address a dearth of workers prepared for science, technology, engineering, and mathematics (STEM) jobs in the USA, K-12 mathematics and science education is increasingly at the center of discussions about how to ensure international competitiveness. For instance, the current presidential administration has launched an “Educate to Innovate” campaign (The White House, 2012), designed to improve the coordination and facilitation of efforts to improve STEM education and prepare the students of today for the jobs of tomorrow.

In the STEM fields, mathematics and science education provide the foundation for advanced knowledge and professional skills that will prepare our nation’s youth to compete for the surge of high-level jobs in engineering and technology (National

---

B. Clarke (✉) · N. Nelson · L. Shanley  
Center on Teaching and Learning, University of Oregon, Eugene TX, USA  
e-mail: clarkeb@uoregon.edu

N. Nelson  
e-mail: nnelson3@uoregon.edu

L. Shanley  
e-mail: shanley2@uoregon.edu

Math Advisory Panel [NMAP], 2008). Arguably, students' understanding of mathematics, starting at an early age, is at the core of their ability to gain access to STEM jobs. Accordingly, proficiency in mathematics is receiving increasing attention, beginning in the early years of a student's education, because the early elementary years represent a critical first step in building a long-term foundation for success in mathematics. Emerging evidence suggests the long-term consequences of struggling early in mathematics exact the same or greater deleterious toll as early reading difficulties (Duncan et al., 2007; Morgan, Farkas, & Wu, 2009). For instance, students struggling to learn mathematics are ill-prepared for well-paying jobs in a modern, technological economy (National Academy of Sciences, 2007). Disparities in mathematical competency are evident between students from different racial and socioeconomic subgroups, impacting the life opportunities of a substantial portion of the population (Siegler et al., 2010). Moreover, mathematics difficulties are as persistent and difficult to remediate as reading difficulties (NMAP, 2008). In other words, just as early intervention in reading is critical, prevention of mathematics difficulties and effective early intervention should also be a primary focus of educational research and practice in mathematics.

Unfortunately, mathematics achievement in the USA is lagging. Results of the 2011 National Assessment for Educational Progress (NAEP) indicate that only 40% of fourth graders scored at or above *proficient* in mathematics, and nearly half of all fourth graders with a disability scored *below basic*. The percentage of students that demonstrate proficiency in mathematics also worsens over time (e.g., 35% of the eighth graders scored at or above *proficient* in mathematics in 2011). On international measures of achievement in fourth and eighth grades, the USA ranks ninth and twelfth, respectively, of approximately 50 countries participating in international benchmarking (Trends in International Mathematics and Science Study: TIMSS, 2011b). Although these rankings indicate students in the USA could be performing far worse, we are also failing to prepare students for the level of mathematics they may need, in order to acquire the 62% of American jobs that will require advanced math skills in the coming decade (Hanushek, Peterson, & Woessmann, 2010). Just 6% of the US students scored at the equivalent of the advanced level in mathematics on the Program for International Student Assessment (Organization for Economic Cooperation Development & Programme for International Student Assessment, 2007), while 30 other countries had a larger percentage of students scoring at this level out of 56 total countries that participated in the assessment (Hanushek et al., 2010). In sum, when it comes to ensuring the ability of our youth to successfully compete for jobs in an international marketplace that requires proficiency in mathematics for technological prowess, we are being outcompeted by a number of countries that do not share the same level of resources we possess in the USA (Hanushek et al., 2010; TIMSS, 2011b).

As competitors in an international marketplace, increasingly driven by technological innovation, it is imperative that US students acquire mathematical proficiency. Results from national and international assessments indicate that we, as a nation, have been inadequate in achieving this aim. The rest of this chapter emphasizes on the role of fluency in mathematical proficiency, discusses the types of interventions

that are employed to promote mathematical fluency, and describes the assessment instruments used to measure mathematical fluency across grade levels.

## Why Focus on Mathematics and Mathematical Fluency?

Despite a clear need to focus on mathematics education, the research based on the development of mathematical proficiency pales in comparison to extensive research that has been conducted in the area of reading (Clarke, Gersten, & Newman-Gonchar, 2010). But we can use what we know about the development of reading skills to inform our thinking about mathematics. For instance, there is broad consensus that foundational (e.g., phonological awareness) and higher order skills (e.g., vocabulary and reading comprehension) are critical areas of reading skill development that must be taught in concert. Congruently, mathematics experts agree that conceptual understanding (i.e., understanding mathematical ideas, the way they function, and the contexts where they apply) must be emphasized alongside efforts to teach procedural fluency, in an intertwined manner (NMAP, 2008; National Research Council [NRC], 2001).

There are also parallels between the types of skills that form the basis of understanding in reading in mathematics. We know, for example, that students learning to read must demonstrate phonemic awareness to have a solid understanding of the sounds that comprise language and become strong readers (National Reading Panel [NRP], 2000). In mathematics, to demonstrate proficiency, students must possess early numeracy skills (e.g., numeral identification, understanding one-to-one correspondence, and magnitude comparison) to understand relations between numbers and quantities (NRC, 2001). Although developmental trajectories in mathematics are often considered more linear (i.e., more advanced skills build directly upon basic skills over time) than the trajectories described for reading development (e.g., students apply similar reading skills in each grade to different types of texts that increase in difficulty as students make progress), the parallels between reading and mathematics in the types of skills and the need to simultaneously emphasize foundational and higher order thinking can inform efforts to improve mathematical proficiency.

Perhaps because research about the development of mathematical proficiency is relatively nascent, there is also substantially less evidence about effective practices for teaching mathematics when compared to our knowledge about effective practices for teaching reading (NMAP, 2008). However, we can learn from the research that has been conducted on reading instruction and intervention in several ways. First, as a result of No Child Left Behind (NCLB), there is increased emphasis on comprehensive systems of support to assist all students in meeting rigorous standards of achievement by 2014. Research in reading has informed the types of assessments (e.g., screening and progress monitoring) and scaffolded supports (e.g., Tier 2, Tier 3 interventions) that comprise these multitiered systems. Similarly, the Institute of Education Sciences (IES) Practice Guide, *Assisting Students Struggling*

*with Mathematics: Response to Intervention for Elementary and Middle Schools* (Gersten et al., 2009), was written to provide guidance to schools and districts looking to establish Response to Intervention (RTI) systems of support in mathematics, using the best evidence available for interventions and assessments. The IES Practice Guide provides support that an RTI approach may be an effective mechanism for supporting the mathematical proficiency of all students.

As momentum shifts toward building service delivery systems of support and identifying the interventions that work to improve students' mathematical proficiency, increased attention has been given to the content that should comprise these interventions. Research on instruction and intervention in mathematics indicates there are key concepts (akin to the five "big ideas" in reading: Coyne, Zipolo, & Ruby, 2006) that should be targeted to support students' proficiency. These key concepts include a focus on whole number concepts in the elementary grades, and an emphasis on rational numbers beginning in fourth grade to support algebra readiness, and other critical foundations of algebra, including key topics in geometry and measurement (Gersten et al., 2009; NMAP, 2008). A number of states have sought to adopt the Common Core State Standards for Mathematics (CCSS-M, 2010). The CCSS-M are widely vetted standards that rest on the NCTM (2000) process standards (i.e., problem-solving, reasoning and proof, communication, connections, and representation) and the principles outlined by the National Research Council (2001) in their volume, *Adding It Up* (i.e., understanding, computing, applying, reasoning, and engaging). The CCSS-M is built on the consensus of experts that conceptual understanding and procedural fluency are critical constructs within mathematics topics, across grades. Recognizing the importance of fluency, one of the eight recommendations in the IES Practice Guide is to "devote about 10 min in each session to building fluent retrieval of basic arithmetic facts" in interventions at all grade levels (Gersten et al., 2009). That is, across grades, experts indicate a need for students to develop automaticity with whole and rational number operations.

Not surprisingly, the bulk of this chapter focuses on the importance of fluency in mathematics; however, we are not advocating that "fluency" is promoted at the cost of conceptual understanding, nor that fluency carries a narrow definition. In fact, we agree with the NCTM (2000) that "developing fluency requires a balance and connection between conceptual understanding and computational proficiency" (p. 35). In addition, we describe how mathematical fluency supports mathematical proficiency for students with learning disabilities (LD) and their typically developing peers, in terms of working memory demands and cognitive load theory.

### ***What is Mathematical Fluency?***

There is overwhelming support from cognitive scientists, researchers, and educators alike that fluency in mathematics supports mathematical proficiency, and should be a focus in Grades K–12 (e.g., NCTM, 2000; NMAP, 2008; NRC, 2001). Traditionally, fluency has been defined in terms of computational proficiency, or being able to quickly and accurately recall basic math facts and procedures. However, this narrow

definition does not take into account the relation between conceptual understanding, procedural knowledge, and basic fact recall, and the notion that demonstrating mathematical fluency requires an awareness of these interconnections. Baroody (2011) defines fluency as the quick, accurate recall of facts and procedures, *and the ability to use them efficiently*. That is, as students develop procedural fluency, it is essential that mastery be tied to conceptual understanding to promote adaptive expertise. In other words, students need to know when they can use an algorithm and when they cannot, in order to demonstrate mathematical fluency. We contend, as do others (e.g., Fennell, 2011) that fluency is a broad construct, which refers to proficiency across mathematical domains (e.g., early numeracy, whole number concepts, rational number concepts, and algebra).

### ***How Does Mathematical Fluency Support Mathematical Proficiency?***

Mathematical fluency provides access to mathematical proficiency through several hypothesized mechanisms. As evidenced by the results of national assessments, students with LD tend to struggle in mathematics to a greater degree than their nondisabled peers (e.g., 2011 NAEP results). Research demonstrates that students with LD in mathematics typically struggle to attain fluency with basic number combinations and simultaneously demonstrate working memory deficits that may be contributing to these “developmental differences” in computational proficiency (Geary, 1996; Geary, Hoard, Byrd-Craven, Nugent, & Numtee, 2007). Students who struggle to automatically retrieve basic number combinations often work more slowly and tend to be more error prone when attempting more complex mathematical problems (Geary, 2004; Jordan & Montani, 1997). Furthermore, fluent basic number combination retrieval has been linked to successful word problem completion, presumably due to reduced working memory demands (Geary & Widaman, 1992; Geary, 2004). For the 5–8% of the students with LD in mathematics, it appears working memory deficits may be inhibiting mathematical fluency, and contributing to generalized difficulties in developing mathematical proficiency.

Working memory may also play a broader role in mathematical proficiency for a range of learners, where students who score lower on a range of working memory tasks demonstrate increased difficulty in mathematics (Raghubar, Barnes, & Hecht, 2010). Several studies have demonstrated that working memory skills predict mathematical fluency and problem-solving, even when controlling for cognitive variables, including attention, intelligence, and phonological processing (Fuchs et al., 2005; Swanson & Beebe-Frankenberger, 2004). In addition, research indicates a range of factors (e.g., age, language, and math representations) may interact with working memory to predict mathematical skill, including mathematical fluency (Raghubar et al., 2010).

There is also support that fluency in mathematics frees cognitive resources for higher order reasoning activities. In their seminal article on reading fluency,

LaBerge and Samuels (1974) argued that human beings can only actively attend to only one thing at a time, thus learners can only do more than one thing at a time if one of the tasks can be performed automatically. Although initially applied to issues of reading performance, these conclusions are relevant to a discussion of the role of fluency in mathematics performance, as well. Advocates for mathematics interventions that train students to become more fluent at targeted mathematics tasks posit that being fluent with mathematics tasks reduces the learner's cognitive load and frees cognitive resources for more complex tasks (Geary, 2004; Geary & Widaman, 1992; Jordan & Montani, 1997).

Broadly, because foundational knowledge and skills unlock the door for understanding of higher order concepts, students who struggle to develop mathematical fluency will struggle to demonstrate mathematical proficiency across their schooling years, with the normative gap growing over time. Take, for instance, the student that is slow and methodical in performing math procedures and recalling number combinations. In elementary school, this student may simply require more time to complete instructional activities. However, when this student encounters a course in Algebra in middle or high school, she may have difficulty understanding daily lessons, because she cannot keep up with the pace of instruction (e.g., even though she understands the procedures, working memory deficits may be preventing access to new concepts). Alternatively, the student may struggle to learn new concepts because she is exhausting cognitive resources solving algorithms (e.g., the cognitive demand of both solving algorithms and learning new algebra concepts is overwhelming in combination). Regardless of the source of the deficit, it is clear that students who are unable to demonstrate fluency in mathematics will fall farther behind their peers who do not struggle with mathematical fluency. Mathematical fluency is, thus, a key ingredient for mathematical proficiency, achievement, and, ultimately, access to life opportunities.

## **Fluency-Based Interventions in Mathematics**

Although mathematical fluency is a key skill for successful mathematics achievement, general mathematics interventions often do not focus primarily on fluency. Instead, general mathematics interventions tend to target concrete mathematical knowledge and skills such as number sense, algorithms, vocabulary, and proofs. However, as noted in the previous section, mathematics fluency interventions that train students to become more fluent at targeted mathematics tasks are important because possessing mathematical fluency frees students' cognitive resources for more complex tasks. If students struggle to automatically retrieve basic number combinations, they will work more slowly and make more errors when solving complex mathematics problems, whereas students who are fluent in basic number combination retrieval are able to complete word problems more accurately.

In school-based settings, versions or elements of fluency interventions are often implemented class wide in the elementary grades as students are expected to master



all basic number combinations to 100. In Grades 6–12, fluency interventions are typically utilized in small group settings as part of specialized academic programs. When implementing targeted mathematics fluency interventions, interventionists aim to improve students' cognitive processes and resources that underlie fluent mathematical performance. However, it can be difficult to determine whether or not improved mathematical performance as measured by fluency assessments discussed in the following section is truly an indicator of cognitive development and not simply an artifact of rote memorization. Ultimately, all fluency interventions operate on the premise that improving mathematical fluency is fundamental to overall improvement in mathematics performance. Mathematics fluency interventions can be categorized into three general types: (1) those that utilize repeated trials with multiple forms to train students to become more fluent at a specific task; (2) interventions that target underlining academic and cognitive skills to teach students generalizable strategies that result in improved mathematics fluency; and (3) general mathematics interventions that include fluency skill-building components to impact both basic number combination proficiency and conceptual fluency.

In the following section, we describe each type of intervention, provide examples of interventions within each category that have been used in research and practice, and summarize research that has been conducted to evaluate the effectiveness of each type of intervention. We follow with a discussion of the challenges related to evaluating the generalizability of mathematics fluency interventions and conclude with summative recommendations to consider when selecting a mathematics fluency intervention.

### ***Repeated Trials Fluency Training***

For many years, researchers and educators have advocated for the use of repeated daily timed mathematics activities to build fact fluency with elementary students by implementing a variety of training components to build rate and accuracy (Miller & Heward, 1992). As repeated practice is a key feature of fluency interventions, most protocol-based fluency training programs rely on discrete learning trials with numerous practice opportunities of the same mathematical material to build speed and accuracy in responding. Some sample programs are detailed in Table 3.1.

Much of the research base for these interventions originates in special education literature and utilizes single-case designs to isolate specific learning gains. Fluency-based interventions tend to target repeated measures of basic number combinations with elementary-aged students as the focal population. As these interventions are tested with a small number of learners, computing effect sizes and making generalized claims about the research findings can be challenging. Single-subject researchers compute effect sizes using techniques that compare the distinct characteristics of student performance in each phase of the study (e.g., pre- and post-intervention). By comparing data points across phases, researchers generate either percentage of nonoverlapping data (PND), interpreted as a percentage with values

**Table 3.1** Sample repeated trial fluency-training interventions

Intervention	Description
Cover, copy, and compare (Skinner, Turco, Beatty, & Rasavage, 1989)	Five step process: (1) look at a model of the math fact with the answer included, (2) cover the math fact with the answer, (3) write the fact with the answer, (4) uncover the original math fact with the answer, and (5) compare
Incremental rehearsal (Burns, 2005)	A flashcard-based drill procedure that combines unknown facts with known facts
Taped-problems (McCallum, Skinner, & Hutchins, 2004)	Using a list of problems on a sheet of paper, the learner is instructed to answer each problem before the answer is provided by an audiotape player using various time delay procedures to adjust the intervals between the problem and answer (adapted from Freeman & McLaughlin's (1984) taped-words intervention)
Detect, practice, and repair (Poncy, Skinner, & O'Mara, 2006)	Multicomponent intervention: (1) metronome-paced, group assessment administered to identify unknown facts, (2) cover, copy, and compare procedures used with unknown facts, (3) 1-min speed drill, and (4) learners graph their accuracy
Math to mastery (Doggett, Henington, & Johnson-Gros, 2006)	Multicomponent intervention: (1) preview problems, (2) repeated practice, (3) immediate corrective feedback, (4) summative and formative feedback, and (5) self-monitoring of progress
Great leaps math (Mercer, Mercer, & Campbell, 2002)	Multistage strategy: (1) greeting and set behavior expectations, (2) review previous facts and progress graph, (3) conduct instructional session with short-timed practice, error correction, and teaching, (4) administer a 1-min fluency probe, (5) graph accuracy

larger than 70% considered meaningful, or a metric called percentage of all non-overlapping data (PAND) and convert this value to a Phi coefficient ( $\phi$ ) that serves a measure of effect size (Parker & Hagan-Burke, 2007; Parker, Hagan-Burke, & Vannest, 2007). Phi is intended to represent the effect of an intervention and can be interpreted with a rule of thumb where values  $\leq 0.20$  are considered small, 0.21 through 0.79 represent a medium effect, and  $\geq 0.80$  are considered large. It should be noted however, that because  $\phi$  is directly tied to the number of data points collected in each phase of a single subject study, the potential values of  $\phi$  are unbounded and it is not uncommon to generate extremely large values when studies have a large number of nonoverlapping data points.

In recent years, meta-analyses have been conducted to compare the treatment effects of various fluency interventions in mathematics and other academic areas (e.g., Coddling, Burns, & Lukito, 2011; Joseph et al., 2012). By grouping interventions according to the fundamental strategy employed or by the general treatment component utilized (e.g., drill, practice with modeling, and self-management), researchers have been able to compare categories of interventions. Perhaps not surprisingly, meta-analytic findings suggest that interventions employing flashcard-based drill activities (e.g., incremental rehearsal) and practice sessions with a modeling component (e.g., math to mastery and great leaps) have proven most effective, with mean  $\phi$  values of 92.00 (extremely high) and 0.71 (moderately

strong), respectively. Self-management strategies that require learners to monitor their own understanding (e.g., cover, copy, and compare) demonstrated moderate effect sizes (mean  $\varphi=0.55$  and mean PND=60.2–70.7) proving productive as well (Coddling et al., 2011; Joseph et al., 2012). However, fluency interventions that prescribed learner practice without a modeling component had little to no impact on student performance (mean  $\varphi=-0.003$ ).

When evaluating additional characteristics of fluency interventions, meta-analytic results suggested that fluency approaches including multiple components with combinations of rehearsal, correction, and practice strategies demonstrated better learner outcomes. Specifically, interventions with more than three components had a moderately strong effect size (mean  $\varphi=0.68$ ) and those with less than three components had a negligible mean  $\varphi$  value (Coddling et al., 2011). Additionally, coupling mathematics fluency interventions based on self-management strategies with other instructional components was found to be effective across numerous studies, mean PND=87.9–97.5 (Joseph et al., 2012).

In addition to conventionally delivered fluency interventions, technology-delivered fluency interventions have become increasingly popular and prolific. Traditionally, computer-aided interventions utilized drill-based procedures providing repeated practice of basic number combinations, but technological advances have allowed intervention developers to incorporate a variety of effective practice and self-management strategies into technology-delivered fluency programs. Programs that present sets of basic number combinations from a specified numerical range (e.g., *flash card program* and *Math Blaster*) are freely available for download and have been used in research programs to compare their utility to peer tutoring and other drill-based procedures (Cates, 2005; Mautone, DuPaul, & Jitendra, 2005). These studies have generated mixed results, with some students responding well to technology-based interventions and others performing better in traditional intervention conditions. Studies of both downloadable basic number combinations programs and researcher-developed mathematics drill programs such as *Math Facts in a Flash* (Renaissance Learning, 2003) have dedicated particular attention to at-risk students for mathematics difficulties. Results of this research suggest that computer-based interventions may result in not only improved mathematical fluency, but also increased on-task behavior (Mautone et al., 2005; Burns, Kanive, & DeGrande, 2012).

When comparing fluency interventions and evaluating their effectiveness for specific populations, it may be that distinct learner characteristics are predictive of the likelihood of responding well to a particular intervention. Research in this area has found that initial level of mathematics fluency can be a significant predictor of intervention effectiveness (Coddling et al., 2007), and meta-analytic findings have suggested that baseline levels of fluency (instructional or frustration) may be associated with differential intervention effectiveness when comparing interventions that either (a) aim to support basic number combination acquisition (acquisition), or (b) intend to bolster learner fluency with known facts (rehearsal) (Burns, Coddling, Boice, & Lukito, 2010). More specifically, the results of this study suggested that initial fluency performance was significantly linked to intervention outcomes such

that acquisition interventions were more effective for learners with a frustration baseline fluency level (mean  $\varphi=0.84$ ) compared to learners with an instructional baseline fluency level (mean  $\varphi=0.49$ ). These findings provide support for the argument that effective mathematics fluency interventions should be implemented with careful consideration of initial learner performance, and also suggest that one should consider the phases of mathematical fluency (e.g., acquisition or rehearsal) when selecting a fluency intervention.

### ***Targeting Generalizable Skills and Behaviors***

Although initial level of fluency is a logical predictor of a learner's response to a repeated trial fluency-training intervention, research has shown that a variety of additional cognitive and behavioral factors are also predictive of both mathematics fluency and general mathematics achievement (Geary, Hoard, Nugent, & Bailey, 2013). Based on these correlational findings, mathematics fluency intervention developers have created and studied programs that target learners' underlying cognitive traits and behavioral tendencies. Rather than directly training learners with repeated trials and regular exposure to basic number combinations, these interventions use mathematics fluency probes primarily as outcome measures and attempt to strengthen the learners' foundational skills by teaching generalizable strategies.

Advocates for generalizable skill (e.g., self-management, goal setting, self-evaluation) interventions argue that teaching students to utilize their cognitive resources more efficiently and effectively will not only translate into improved fluency, but improved general mathematics achievement as well. Research on behavioral self-management interventions has suggested that these strategies can improve both mathematics fluency and academic engagement, and generalize to more complex mathematical tasks (McDougall & Brady, 1998; Farrell & McDougall, 2008). Performance feedback and goal setting have also been studied as mathematics fluency interventions. Results from these studies have indicated that there is an association between goal setting and feedback-based interventions and improved performance on mathematics fluency measures (Coddington, 2003; Figarola et al., 2008). The challenge in evaluating these interventions is that it can be difficult to isolate the link between improved fluency and the underlying cognitive and behavioral factors. As these interventions rely on the repeated administration of fluency probes to monitor student progress, one could argue that mathematics fluency improvements could simply be due to the additional fluency practice and residual testing effects resulting from regular fluency probe administration.

### ***Mathematics Interventions with Fluency Skill-Building Components***

Rather than targeting underlying skills through cognitive training intended to support performance on both basic and complex mathematical tasks, others advocate

for allocating intervention resources to boost general mathematical knowledge based on its relation with both accuracy in basic number combinations and general mathematics skill development. For example, because number sense performance in kindergarten can predict later calculation fluency above and beyond cognitive factors (Locuniak & Jordan, 2008), researchers claim that early academic interventions support the acquisition of foundational skills that are pre- or corequisites of mathematics fluency. In addition to boosting the development of foundational academic skills, many general mathematics interventions include fluency-training components to build speed and accuracy with targeted mathematical material. In fact, researchers recommend that mathematics interventions include fluency exercises (Fuchs et al., 2008a; Gersten, Jordan, & Flojo, 2005), and there is a high prevalence of fluency components in successful mathematics intervention curricula (Bryant et al., 2008; Fuchs, Fuchs, & Hollenbeck, 2007; Ketterlin-Geller, Chard, & Fien, 2008; Jitendra et al., 2013). Results from intervention research conducted by Fuchs et al. (2008) suggested that efforts to improve general mathematics skills and performance on complex mathematical tasks should be supported by mathematical fluency skill building. Although, improving fluency is not the primary objective of most general mathematics interventions, computational fluency is considered an essential aspect of mathematical performance and often explicitly addressed in intervention curricula aimed at at-risk students.

### *Challenges in Establishing Generalizable Interventions*

Although some have argued that fluency is an interwoven component of applied problem-solving (Lin & Kubina, 2005) and improved fluency is associated with improved performance on more complex tasks (VanDerHeyden & Burns, 2009), others have found that fluency does not generalize across mathematics problems or skills (Poncy, Duhon, Lee, & Key, 2010). Poncy and colleagues suggest that fluency instruction targeting basic declarative skills (i.e., basic number combinations) needs to be supplemented with instruction that supports the fluent completion of procedural, multistep tasks for fluency to generalize to overall mathematics performance. In sum, general research evidence suggests that for the mathematics fluency interventions to be optimally effective, they should utilize a variety of strategies to train learners to be more fluent with basic number combinations and be integrated into the general mathematics instructional program to support skill transfer and generalization.

The variety of mathematics fluency intervention approaches speaks to the lingering debate about the generalizability of fluency and the role of automaticity with foundational material in facilitating advanced mathematical achievement. The debate about the role of fluency in mathematics parallels similar debates about the nature of the relation between fluency and comprehension in reading. Few years ago, Slocum, Street, and Gilbert (1995) found that interventions that proved effective at increasing reading rate had unreliable impacts on reading comprehension. They also noted challenges related to (a) identifying sensitive outcome measures of

general reading performance and (b) the experimental design of the study when attempting to examine the mechanisms that link fluency and general reading achievement. Similar challenges abound in mathematics fluency research. Additional research is needed to investigate the mechanisms that link mathematical fluency and overall mathematics performance and determine how one can isolate intervention techniques that target rate of responding (considered true fluency) from repeated exposure or additional practice, two common features of fluency interventions that can increase overall mathematics performance on their own regardless of whether or not the interventions improve general fluency proficiency (Doughty, Chase, & O'Shields, 2004). The effect of mathematics fluency training interventions is evaluated by comparing pre- and posttests of student performance on basic number combination probes, but it can be difficult to isolate the source of those gains. Improved performance on fluency probes is often assumed to be evidence of improved rate of responding, but could also be the result of increased knowledge or the simple acquisition of basic number combinations alone. Effective assessments of mathematical fluency are critical to identifying factors of effective interventions and simply measuring student progress. In the next section, we will examine how mathematical fluency is measured and the role that fluency plays in mathematics assessment.

## **Fluency and Mathematics Assessment**

The relation between fluency and mathematics assessment is complex. At first glance, the complexity of this relation is not readily apparent. In simple terms, a large number of commonly used mathematics assessments are timed, and a timed measure seems to imply that the measure functions as a fluency measure. However, a more in-depth examination of commonly used mathematics measures reveals a more dynamic relation between the construct of fluency and mathematics assessment.

To fully explore the role of fluency in mathematics assessment, we first examine the original development of widely used measures that are considered to be fluency-based mathematics assessments and their intended use in educational decision-making. We follow by providing an overview of measures currently in use and conclude with a discussion examining critical unanswered questions to which we feel the field should be attuned as we attempt to advance in both research and practice.

### ***How Are We Measuring Fluency?***

The construct of fluency in mathematics assessment is typically examined within the realm of a set of measures broadly classified as curriculum-based measures or CBM. Math CBM (M-CBM) measures have a long history and the general CBM category includes an expanding set of instruments used for a variety of purposes



by schools such as screening, program evaluation, and monitoring student growth (Deno, 2003; Deno & Mirkin, 1977). Originally M-CBM measures focused on a student's understanding of computation objectives and application of conceptual understanding to problem-solving for the elementary school grades. But now the umbrella of M-CBM measures includes an array of measures designed to cover student development in mathematics from beginning number sense in the early elementary grades (Gersten et al., 2012) to a student's understanding of pre-algebra in middle school (Foegen, 2008). Across this spectrum, content-assessed ranges from students comparing the magnitude of two one-digit numbers to combining like integers. Yet across this vast range of mathematics content one central feature remains prevalent—a timing element. But why is a timing element a common universal feature of almost all M-CBM measures?

The design of M-CBM measures was governed by a multitude of considerations including the content assessed and technical characteristics (Deno, 2003). But serious consideration was also given toward the practical application of their use in schools. Because the original intent of CBM measures was to monitor the growth of at-risk students to gauge their response to instructional interventions and modifications (i.e., progress monitoring), the measures needed to have certain design characteristics that enabled them to be administered frequently and repeatedly over time (Deno, 1985). It was this consideration that played a major role in the inclusion of a timing element. Seminal articles detailing the use and design features of the measures were linked to their need to be used in a repeated fashion and the importance of efficient measures to meet that goal.

Typically, an M-CBM battery consists of two measures; a computation measure that covers major topics in the standards relating to computation, and a concepts and applications subtest that assesses all other topics including word problems, measurement, money and time, and geometry. While the computation and concepts and applications approach to M-CBM measures has long been utilized, a new theoretical framework has been advocated and initially researched that explores the possibility that math disabilities can occur in one of the two areas or both simultaneously (Fuchs, Fuchs, & Zumeta, 2008). M-CBM measures demonstrate acceptable test-retest, inter-rater and alternate-form reliability, and concurrent and predictive validity between .50 and .60 (Foegen, Jiban, & Deno, 2007). The timing of the measures varies by grade level with shorter durations (1 or 2 min) in the earlier grades, and up to 5 min for the later grades.

Although originally M-CBM measures were designed to align with actual curricula (i.e., the C in CBM stood for a specific curriculum) over time new iterations of M-CBM measures were designed to align to specific state standards (Gersten et al., 2012) and other similar but non-timed measures were aligned to foundational documents such as the National Council of Teachers of Mathematics Focal Points (2006; Clarke et al., 2011). This trend has specific implications for future measurement development as more contemporary standards, such as the Common Core, are adopted and implemented. Other advancements in the use of M-CBM have focused on extending the use of M-CBM-like measures to the early elementary and middle school grades. In the next section, we detail developments in those age and grade ranges.



## Fluency-Based Measures Assessing Number Sense

At the early elementary grades (kindergarten and first grade) fluency-based measures are designed to tap into a student's beginning and developing number sense. Although the concept of number sense is widely accepted it has been elusive to operationalize. It has been postulated as a corollary to phonological awareness and described by Gersten and Chard (1999) as "a child's fluidity and flexibility with numbers, the sense of what numbers mean, and an ability to perform mental mathematics and look at the world and make comparisons" (p. 19). Other researchers have noted the complexity of attempting to define number sense but at the same time attempted to begin articulating exactly what is number sense (Berch, 2005).

Possessing number sense ostensibly permits one to achieve everything from understanding the meaning of numbers to developing strategies for solving complex math problems; from making simple magnitude comparisons to inventing procedures for conducting numerical operations; and from recognizing gross numerical errors to using quantitative methods for communicating, processing, and interpreting information. (p. 334).

The complexity in defining number sense is often encapsulated by the wide range of specific number proficiencies put forth as indicating an underlying understanding of number. That is, although there is a general consensus on what number sense is, the specific proficiencies that capture number sense are varied. The National Research Council's (2009) *Mathematics Learning in Early Childhood* recognized the inherent difficulty in operationally defining number sense and noted that any attempt to measure number sense would likely focus on assessing key proficiencies (e.g., applying number properties or counting strategies to solving addition and subtraction problems and simple word problems). Thus, while measures developed to assess number sense would assess specific proficiencies, the larger goal was for the measure to tap into the underlying construct of number sense. Despite the complexity and difficulty in measuring number sense through examining specific skills, a number of assessments have been developed. Typically, these assessments focus on key constructs of beginning number sense.

In the next section, we detail and summarize that work focusing on three components of number sense judged to be critical by cognitive psychologists and education researchers: magnitude comparison (Booth & Siegler, 2006), strategic counting (Geary, 2004), and basic fact fluency (Jordan, Hanich, & Kaplan, 2003). The overview is not intended to suggest other aspects of mathematics development and number sense are not critical (e.g., solving word problems) or to suggest that other timed measures have not been developed to assess number sense or math readiness (e.g., numeral identification) but rather to focus on those measures tapping critical constructs and do so in a manner that is focused on a student's fluency with the construct. It should also be noted that although the measures and constructs reviewed focus on a specific skill, the original development of the measures mirrors that of early CBM development in that the goal is to provide a powerful indicator of a student's broader understanding of the domain. Thus, while a measure may have a student complete a specific number sense or mathematics task (e.g., noting the missing number in a sequence of numbers) the measures are intended to provide an indicator or overall level of understanding.

## Magnitude Comparison

Magnitude comparison is made up of a number of specific skills but fundamentally it is based on the ability to draw comparisons about relative magnitude. Magnitude comparison can include the ability to determine which number is the greatest in a set and to be able to weigh relative differences in magnitude quickly and accurately. For example, initially children may know that 5 is bigger than 2 and then begin to understand that 7 is also bigger than 2 and that the difference between 7 and 2 is greater than the difference between 5 and 2. As children advance to developing a more nuanced understanding of number and quantity, they are able to make increasingly complex judgments about magnitude. In the earlier grades, the development of an understanding of magnitude is a critical underpinning of the ability to calculate.

It has been hypothesized that as children develop a greater understanding of magnitude, they map that understanding onto a mental number line and begin to use that mental number line to further understand magnitude and to solve initial calculation problems (Dehaene, 1997). For example, when a student is presented a problem to add 4 and 2, a student who can recognize 4 as the greater magnitude can then solve the problem by counting up 2 (this example also implies an understanding of the commutative property and the use of strategic counting) on a mental number line to derive a correct answer.

Typically, measures of magnitude comparison require a student to identify the greater number from a set of two numbers. A number of research teams have designed and tested similar measures of magnitude comparison for kindergarten and first grade with all measures including a timing element but varying the range of numbers used in the materials in response to potential concerns about floor or ceiling effects. For example, some measures use number sets from 0 to 10 for kindergartners (Lembke & Foegen, 2009; Seethaler & Fuchs, 2010), while others use 0–20 (Clarke et al., 2011).

A recent overview of screening measures in the early grades (Gersten et al., 2012), noted strong reliability coefficients across studies of examining magnitude comparison measures. Evaluations included interscorer, alternate-form, and test–retest, all of which reported coefficients consistently greater than .80, and concurrent and predictive validity data correlating with summative measures of mathematics falling mostly in the .50–.70 range.

## Strategic Counting

Strategic counting is fundamental to developing mathematical understanding and proficiency and has been defined as the ability to understand how to count efficiently and to employ efficient counting strategies to solve an array of problems (Siegler & Robinson, 1982). Students who fail to develop strategic counting and to utilize counting principles efficiently to solve problems are more likely to be classified as having a mathematics learning disability (Geary, 1994). As children develop strategic counting strategies they are more able to efficiently solve addition and subtraction problems by applying this knowledge in combination with a

growing understanding of number properties. For example, a child who understands counting up (e.g.,  $5+2$  can be solved by counting up from 5) and the commutative property (i.e.,  $a+b=b+a$ ) can apply the min strategy (counting up from the larger addend) so if given a problem “what is 6 more than 3?” she will solve the problem by changing the problem to “what is 3 more than 6?” and simply count on from 6 to derive the answer.

The most common strategic counting measures require students to determine the missing number from a sequence of numbers. Similar to magnitude comparison measures, strategic counting measures include a timing element and vary the range of numbers used based on the grade level to avoid floor or ceiling effects. Some researchers have begun to experiment with measures that require skip counting (e.g., filling in the blank in a number series, 5, 10, \_\_, 20) (Lembke & Foegen, 2009) An overview of strategic counting measures found moderate concurrent and predictive validities (range = .37–.72) and strong reliabilities (range from .59 to .98) (Gersten et al., 2012).

### Retrieval of Basic Arithmetic Facts

An established finding in the research based on mathematics disabilities has been that students who are diagnosed as mathematics LD exhibit consistent and persistent deficits with the automatic retrieval of addition and subtraction number combinations (Goldman, Pellegrino, & Mertz, 1988; Hasselbring et al., 1987). Geary (2004) found that children with difficulties in mathematics typically fail to make the transformation from using simple strategies to solve problems (e.g., by counting on their fingers or with objects) to solving problems mentally without using these objects (also Jordan, Kaplan, Ramineni, & Locuniak, 2009).

Research trends seem to indicate that, although students with mathematics LD often make progress in their use of algorithms when provided with classroom instruction, significant deficits remain in their ability to retrieve basic number combinations (Geary, 2004; 2001; Jordan et al., 2003). A number of theories have been put forth to explain these difficulties. Geary (2004) hypothesized that the difficulty was related to issues with semantic memory (i.e., the ability to store and retrieve abstract information efficiently). Jordan et al. (2003) hypothesized that fact-retrieval difficulty was rooted in weak number sense, and that when students lack number sense and an understanding of the relations between and among numbers and operations they fail to develop automaticity with addition and subtraction number combinations. Whatever the root cause of difficulty with addition and subtraction number combinations, they remain a powerful predictor of later mathematics achievement (Jordan et al., 2009). Initial research on number combination or fact fluency measures shows promise in the early elementary grades (first and second grade; Bryant, Bryant, Gersten, Scammacca, & Chavez, 2008; Gersten, Clarke, Dimino, & Rolffhus, 2010).

## Fluency-Based Measures in the Middle School Grades

As students advance to the middle school grades, new CBM-like measures have been designed to assess critical concepts of algebra (Foegen, 2008), problem-solving (Montague, Penfield, Enders, & Huang, 2010), and estimation (Foegen & Deno, 2001). Similar to M-CBM computation and concepts, and applications measures for the same grade range, these new CBM-like measures provide more time (e.g., 5 min) for students to work. The complexity of mathematics skills assessed by upper-grade measures brings into question how well we can assess mathematics using a timed measure. Consider one of the algebra measures developed by Foegen (2008) designed to assess, among other features, the following basic skills in algebra: applying the distributive property, working with integers, combining like terms, and simplifying equations. Whether or not a timed measure (and of what duration) is the best approach to assessing this content is a legitimate question along with considering how untimed or measures with a longer duration fit into different types of assessments (e.g., screening and progress monitoring).

In part, the issue of timing represents the larger issue of whether or not a timed measure is also a fluency measure. Given that the original purpose of developing CBM measures was to provide an initial gauge of student understanding in a topic and a long-term analysis of growth in that topic, one could argue that not all of the measures reviewed in this chapter are fluency measures. However, given that all the measures do assess how quickly and accurately a student applies specific skills (whether for 1 min or for 5), they do assess fluency. The answer likely lies between those two positions in that the measures provide useful information in both providing an indicator of overall student understanding and a student's fluency with greater overlap between the two in the earlier grades.

## Conclusion

The concept of fluency and its importance is well established and accepted in the field of mathematics. Seminal documents on mathematics instruction readily acknowledge the role of fluency in the development of student proficiency in mathematics (NMAP, 2008; Gersten et al., 2009). Perhaps in a proactive attempt to avoid the “reading wars” that have plagued the field of reading instruction, the mathematics field has been more overt and proactive in advocating for viewing fluency in conjunction with the development of conceptual understanding (NMAP, 2008). That is, fluency and conceptual understanding are both of importance and that growth in one fuels increased growth in the other rather than one aspect of mathematics being developed at the cost of another (Wu, 2005).

Given the general acceptance of fluency's importance, continued evaluation of existing and development of new interventions specifically designed to impact flu-

ency seems likely. We consider the development and research efforts reviewed in this chapter as a solid foundation for further work. We believe going forward two important considerations should guide the field. First, if researchers provide only a fluency intervention and evaluate the impact of that intervention with a measure that is closely aligned to the intervention, caution should be exercised when interpreting results. In particular if that measure is considered by the field to provide an overall index of understanding in the broader domain of mathematics. For example, an intervention may focus exclusively on building fluency in identifying the greater of a set of numbers and use a measure of magnitude comparison to examine impact. But because the intervention is specifically targeted on magnitude comparison, increased scores on a measure of the same content may not reflect a generalized improvement in the underlying domain of number sense. Second, given the high probability that low levels in fluency are accompanied by deficits in other areas of mathematics, fluency interventions should rarely be delivered in isolation. That is, students who struggle with fluency in mathematics need a comprehensive intervention that includes, but is not limited to, addressing fluency-related problems. This position is not to say that isolated intervention and research conducted to date lacks importance, it is rather to acknowledge that students with severe deficits in mathematics need an intervention of an intensity equal to their deficits and that likely involves a sustained effort to build conceptual understanding of critical mathematics concepts.

Lastly, developers of current and future measures of mathematics that include a timing element should be proactive in laying forth what constructs they are measuring and how they view the development and use of their measures. A cautionary tale from reading illustrates the point. When Reading First advanced the framework of five big ideas of beginning reading instruction, including accuracy and fluency with connected text (Baker, Fien, & Baker, 2010), states and districts viewed this framework as specifying a need to measure each big idea. The previous role of oral reading fluency as a measure of overall reading health was to some extent replaced with oral reading fluency serving only as a measure of accuracy and fluency with connected text despite the continued evidence that oral reading fluency continues to be validated as a strong measure of general reading achievement including comprehension (Fuchs et al., 2001). Thus, if developers and researchers design and view their mathematics measures as assessing student understanding in a broader domain but a timing element is also included, they should be proactive in discussing and demonstrating the link between their measure and greater understanding in mathematics.

We believe that efforts in all of these areas will help further our understanding of the role of fluency in developing mathematics proficiency. As we advance our understanding, we believe that the field will be better positioned to ensure that all children achieve success in mathematics.

## References

- Baker, S.K., Fien, H., & Baker, D. L. (2010). Robust reading instruction in the early grades: Conceptual and practical issues in the integration and evaluation of tier 1 and tier 2 instructional supports. *Focus on Exceptional Children*, 3(9), 1–21.
- Baroody, A. J. (2011). Learning: A framework. In F. Fennell (Ed.), *Achieving fluency: Special education and mathematics* (pp. 15–58). Reston: National Council of Teachers of Mathematics.
- Berch, D. B. (2005). Making sense of number sense: Implication for children with mathematical disabilities. *Journal of Learning Disabilities*, 38, 333–339. doi:10.1177/00222194050380040901.
- Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, 41, 189–201. doi:10.1037/0012-1649.41.6.189.
- Bryant, D.P., Bryant, B.R., Gersten, R., Scammacca, N., & Chavez, M. (2008a). Mathematics intervention for first and second grade students with mathematics difficulties: The effects of tier 2 intervention delivered as booster lessons. *Remedial and Special Education*, 29(1), 20–32. doi:10.1177/0741932507309712.
- Burns, M. K. (2005). Using incremental rehearsal to increase fluency of single-digit multiplication facts with children identified as learning disabled in mathematics computation. *Education and Treatment of Children*, 28(3), 237–249.
- Burns, M. K., Codding, R. S., Boice, C. H., & Lukito, G. (2010). Meta-analysis of acquisition and fluency math interventions with instructional and frustration level skills: Evidence for a skill-by-treatment interaction. *School Psychology Review*, 39, 69–83.
- Burns, M. K., Kanive, R., & DeGrande, M. (2012). Effect of a computer-delivered math fact intervention as a supplemental intervention for math in third and fourth grades. *Remedial and Special Education*, 33, 184–191. doi:10.1177/0741932510381652.
- Cates, G. L. (2005). Effects of peer versus computer-assisted drill on mathematics response rates. *Psychology in the Schools*, 42, 637–646. doi:10.1002/pits.20105.
- Clarke, B., & Shinn, M. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, 33, 234–248.
- Clarke, B., Baker, S., Smolkowski, K., & Chard, D. J. (2008). An analysis of early numeracy curriculum-based measurement: Examining the role of growth in student outcomes. *Remedial and Special Education*, 29, 46–57. doi:10.1177/0741932507309694.
- Clarke, B., Gersten, R., & Newman-Gonchar, R. (2010). RTI in Mathematics: beginnings of a knowledge base. In S. Vaughn & T. Glover (Eds.), *Advances in response to intervention*. New York: Guilford.
- Clarke, B., Nese, J. F. T., Alonzo, J., Smith, J. L. M., Tindal, G., Kame'enui, E. J., & Baker, S. K. (2011). Classification accuracy of easy CBM first-grade mathematics measures: Findings and implications for the field. *Assessment for Effective Intervention* (Advance online publication). doi:10.1177/1534508411414153.
- Codding, R. S. (2003). *Examining the efficacy of performance feedback and goal-setting interventions in children with AD/HD: A comparison of two methods of goal setting*. (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses A&I database. (Accession No. 305246068).
- Codding, R. S., Shiyko, M., Russo, M., Birch, S., Fanning, E., & Jaspens, D. (2007). Comparing mathematics interventions: Does initial level of fluency predict intervention effectiveness?. *Journal of School Psychology*, 45, 603–617. doi:10.1016/j.jsp.2007.06.005
- Codding, R. S., Burns, M. K., & Lukito, G. (2011). Meta-Analysis of mathematic basic-fact fluency interventions: A component analysis. *Learning Disabilities Research and Practice*, 26(1), 36–47. doi:10.1111/j.1540-5826.2010.00323.x.
- Common Core State Standards Initiative. (2010). *Common core standards for English language arts & literacy in history/social studies, science, and technical subjects*. [http://www.corestandards.org/assets/CCSSI\\_ELA%20Standards.pdf](http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf).



- Coyne, M. D., Zipoli, R. P., & Ruby, M. F. (2006). Beginning reading instruction for students at risk for reading disabilities: What, how, and when. *Intervention in School and Clinic, 41*, 161–168. doi: 10.1177/10534512060410030601
- Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. New York: Oxford University Press.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.
- Deno, S. L. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention, 28*(3–4), 312. doi:10.1177/073724770302800302.
- Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Reston: Council for Exceptional Children.
- Doggett, R. A., Henington, C., & Johnson-Gros, K. N. (2006). *Math to mastery: A direct instruction remedial math intervention designed to increase student fluency with basic math facts*. Unpublished manuscript, Mississippi State University, Mississippi State, MS.
- Doughty, S. S., Chase, P. N., & O'Shields, E. M. (2004). Effects of rate building on fluent performance: A review and commentary. *The Behavior Analyst, 27*, 7–23.
- Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Japel, C., et al. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446.
- Farrell, A., & McDougall, D. (2008). Self-monitoring of pace to improve math fluency of high school students with disabilities. *Behavior Analysis in Practice, 1*(2), 2–35.
- Fennell, F. (2011). All means all. In F. Fennell (Ed.), *Achieving fluency: Special education and mathematics* (pp. 1–14). Reston: National Council of Teachers of Mathematics.
- Figarola, P. M., Gunter, P. L., Reffel, J. M., Worth, S. R., Hummel, J., & Gerber, B. L. (2008). Effects of self-graphing and goal setting on the math fact fluency of students with disabilities. *Behavior Analysis in Practice, 1*, 36–41.
- Foegen, A. (2008). Algebra progress monitoring and interventions for students with learning disabilities. *Learning Disability Quarterly, 31*, 65–78.
- Foegen, A., & Deno, S. (2001). Identifying growth indicators of low-achieving students in middle school mathematics. *Journal of Special Education, 35*(1), 4–16. doi:10.1177/002246690103500102.
- Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in Mathematics. *Journal of Special Education, 41*, 121–139. doi:10.1177/00224669070410020101.
- Freeman, T. J., & McLaughlin, T. F. (1984). Effects of a taped-words treatment procedure on learning disabled students' sight-word oral reading. *Learning Disability Quarterly, 7*, 49–54. doi: 10.2307/1510261
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239–256.
- Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97*, 493–513.
- Fuchs, L. S., Fuchs, D., & Hollenbeck, K. N. (2007). Extending responsiveness to intervention to mathematics at first and third grades. *Learning Disabilities Research and Practice, 22*(1), 13–24. doi:10.1111/j.1540-5826.2007.00227.x
- Fuchs, L. S., Fuchs, D., Powell, S. R., Seethaler, P. M., Cirino, P. T., & Fletcher, J. M. (2008a). Intensive intervention for students with mathematics disabilities: Seven principles of effective practice. *Learning Disability Quarterly: Journal of the Division for Children with Learning Disabilities, 31*, 79–92.
- Fuchs, L.S., Fuchs, D., & Zumeta, R.O. (2008b). Response to intervention: A strategy for the prevention and identification of learning disabilities. In E. L. Grigorenko (Ed.), *Educating individuals with disabilities: IDEIA 2004 and beyond* (pp. 115–135). New York: Springer.
- Geary, D. C. (1994). *Children's mathematical development: Research and practical applications* (1st ed.). Washington, DC: American Psychological Association.



- Geary, D.C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences*, *19*, 229–284.
- Geary, D. C. (2001). Numerical and arithmetical deficits in learning-disabled children: Relation to dyscalculia and dyslexia. *Aphasiology*, *15*, 635–641. doi:10.1080/02687040143000113.
- Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities*, *37*(1), 4–15. doi:10.1177/00222194040370010201.
- Geary, D. C., & Widaman, K. F. (1992). Numerical cognition: On the convergence of componential and psychometric models. *Intelligence*, *16*, 47–80. doi:10.1016/0160-2896(92)90025-M.
- Geary, D. C., Hoard, M. K., Byrd Craven, J., Nugent, L., & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. *Child Development*, *78*, 1343–1359. doi:10.1111/j.1467-8624.2007.01069.x.
- Geary D.C., Hoard M.K., Nugent L., Bailey D.H. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PLoS One*, *8*(1): e54651. doi:10.1371/journal.pone.0054651.
- Gersten, R., & Chard, D. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *The Journal of Special Education*, *33*(1), 18–28. doi:10.1177/002246699903300102.
- Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, *38*, 293–304. doi:10.1177/0022194050380040301.
- Gersten, R. M., Beckmann, S., Clarke, B., Foegen, A., March, L., Star, J. R., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to intervention (RTI) for elementary and middle schools (Practice Guide Report No. NCEE 2009-4060)*. Washington, DC: U.S. Department of Education, National Center for Education Evaluation and Regional Assistance.
- Gersten, R., Clarke, B., Dimino, J., & Rolffhus, E. (2010). *Universal screening measures of number sense and number proficiency for K-1: Preliminary findings (Report No. 2010-1)*. Los Alamitos: Instructional Research Group.
- Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children*, *78*, 423–445.
- Goldman, S. R., Pellegrino, J. W., & Mertz, D. L. (1988). Extended practice of basic addition facts: Strategy changes in learning disabled students. *Cognition and Instruction*, *5*, 223–265. doi:10.1207/s1532690xci0503\_2.
- Hanushek, E. A., Peterson, P. E., & Woessmann, L. (2010). *U.S. math performance in global perspective: How well does each state do at producing high-achieving students?* Cambridge: Program on Education Policy and Governance & Education, Harvard University Kennedy School.
- Hasselbring, T., Sherwood, R., Bransford, J., Fleenor, K., Griffith, D., & Goin, L. (1987). An evaluation of a level-one instructional videodisc program. *Journal of Educational Technology Systems*, *16*, 151–169. doi:10.2190/BR31-J510-CXM4-K41E.
- Jitendra, A. K., Rodriguez, M., Kanive, R., Huang, J. P., Church, C., Corroy, K. A., & Zaslofsky, A. (2013). Impact of small-group tutoring interventions on the mathematical problem solving and achievement of third-grade students with mathematics difficulties. *Learning Disability Quarterly*, *36*(1), 21–35. doi:10.1177/0731948712457561.
- Jordan, N. C., & Montani, T. O. (1997). Cognitive arithmetic and problem solving: A comparison of children with specific and general mathematics difficulties. *Journal of Learning Disabilities*, *30*, 624–634. doi:10.1177/002221949703000606.
- Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology*, *85*, 103–119. doi:10.1016/S0022-0965(03)00032-8.
- Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*, 850–867. doi:10.1037/a0014939.

- Joseph, L. M., Konrad, M., Cates, G., Vajcner, T., Eveleigh, E., & Fishley, K. M. (2012). A meta-analytic review of the cover-copy-compare and variations of this self-management procedure. *Psychology in the Schools, 49*, 122–136. doi:10.1002/pits.20622.
- Ketterlin-Geller, L. R., Chard, D. J., & Fien, H. (2008). Making connections in mathematics conceptual mathematics intervention for low-performing students. *Remedial and Special Education, 29*(1), 33–45. doi:10.1177/0741932507309711.
- LaBerge, D., & Samuels, S.J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293–323. doi:10.1016/0010-0285(74)90015-2.
- Lembke, E. S., & Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and grade 1 students. *Learning Disabilities Research and Practice, 24*, 12–20. doi:10.1111/j.1540-5826.2008.01273.x.
- Lin, F. Y., & Kubina, R. M. (2005). A preliminary investigation of the relationship between fluency and application for multiplication. *Journal of Behavioral Education, 14*, 73–87. doi:10.1007/s10864-005-2703-z.
- Lockard, C. B., & Wolf, M. (2012). Occupational employment projections to 2020. *Monthly Labor Review, 135*, 84–108.
- Locuniak, M. N., & Jordan, N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of Learning Disabilities, 41*, 451–459. doi:10.1177/0022219408321126.
- Mautone, J. A., DuPaul, G. J., & Jitendra, A. K. (2005). The effects of computer-assisted instruction on the mathematics performance and classroom behavior of children with ADHD. *Journal of Attention Disorders, 9*, 301–312. doi:10.1177/1087054705278832.
- McCallum, E., Skinner, C. H., & Hutchins, H. (2004). The taped-problems intervention. *Journal of Applied School Psychology, 20*, 129–147.
- McDougall, D., & Brady, M. P. (1998). Initiating and fading self-management interventions to increase math fluency in general education classes. *Exceptional Children, 64*, 151–166.
- Mercer, C. D., Mercer, K. D., & Campbell, K. U. (2002). *Great leaps math*. Gainesville: Diarmuid.
- Miller, A. D., & Heward, W. L. (1992). Do your students really know their math facts? Using daily time trials to build fluency. *Intervention in School and Clinic, 28*, 98–104.
- Montague, M., Penfield, R. D., Enders, C., & Huang, J. (2010). Curriculum-based measurement of math problem solving: A methodology and rationale for establishing equivalence of scores. *Journal of School Psychology, 48*, 39–52. doi:10.1016/j.jsp.2009.08.002.
- Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities, 42*, 306–321. doi:10.1177/0022219408331037.
- National Academy of Sciences. (2007). *Is America falling off the flat earth?* In N. R. Augustine (Ed.), *Rising Above the Gathering Storm* Committee, National Academy of Sciences, National Academy of Engineering, and Institute of Medicine of the National Academies. Washington, DC: National Academies Press. Retrieved from <http://www.nap.edu/catalog/12021/is-america-falling-off-the-flat-earth>
- National Center for Education Statistics. (2011a). *National Assessment of Educational Progress (NAEP). District of Columbia*: Washington, DC.: U.S. Dept. of Education, Institute of Education Sciences.
- National Center for Education Statistics. (2011b). *Trends in International Mathematics and Science Study (TIMSS). District of Columbia*: Washington, DC: U.S. Dept. of Education, Institute of Education Sciences.
- National Council of Teachers of Mathematics. (2000). *Standards 2000 project*. <http://standards.nctm.org/document/index.htm>.
- National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. <http://www.nctm.org/standards/focal-points.aspx?id=282>.
- National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: US Department of Education.

- National Reading Panel. (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development.
- National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: Mathematics Learning Study Committee.
- National Research Council. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. Washington, DC: National Academies Press. Retrieved from [http://www.nap.edu/catalog.php?record\\_id=12519](http://www.nap.edu/catalog.php?record_id=12519)
- Parker, R. I., & Hagan-Burke, S. (2007). Useful effect size interpretations for single-case research. *Behavior Therapy, 38*, 95–105. doi:10.1016/j.beth.2006.05.002.
- Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education, 40*, 194–204. doi:10.1177/00224669070400040101.
- Poncy, B. C., Skinner, C. H., & O'Mara, T. (2006). Detect, practice, and repair: The effects of a classwide intervention on elementary students' math-fact fluency. *Journal of Evidence-Based Practices for Schools; Journal of Evidence-Based Practices for Schools, 7*, 47–68.
- Poncy, B. C., Duhon, G. J., Lee, S. B., & Key, A. (2010). Evaluation of techniques to promote generalization with basic math fact skills. *Journal of Behavioral Education, 19*(1), 76–92. doi:10.1007/s10864-010-9101-x.
- Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences, 20*(2), 110–122.
- Renaissance Learning. (2003). *Math facts in a flash*. Wisconsin Rapids: Renaissance Learning.
- Seethaler, P. M., & Fuchs, L. S. (2010). The predictive utility of kindergarten screening for math difficulty. *Exceptional Children, 77*, 37–60.
- Siegler, R. S., & Robinson, M. (1982). The development of numerical understandings. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior* (pp. 241–311). New York: Academic.
- Siegler, R., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., Thompson, L., & Wray, J. (2010). *Developing effective fractions instruction for kindergarten through 8th grade: A practice guide (NCEE #2010-4039)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. [whatworks.ed.gov/publications/practiceguides](http://whatworks.ed.gov/publications/practiceguides).
- Skinner, C. H., Turco, T. L., Beatty, K. L., & Rasavage, C. (1989). Cover, copy, and compare: An intervention for increasing multiplication performance. *School Psychology Review, 18*, 212–220.
- Slocum, T. A., Street, E. M., & Gilberts, G. (1995). A review of research and theory on the relation between oral reading rate and reading comprehension. *Journal of Behavioral Education, 5*, 377–398. doi:10.1007/BF02114539.
- Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology, 96*, 471–491.
- The White House. (2012, January). [Official web site of the White House and President Barack Obama]. <http://www.whitehouse.gov/issues/education/k-12/educate-innovate>.
- VanDerHeyden, A. M., & Burns, M. K. (2009). Performance indicators in math: Implications for brief experimental analysis of academic performance. *Journal of Behavioral Education, 18*, 71–91. doi:10.1007/s10864-009-9081-x.
- Wu, H. (2005). *Must content dictate pedagogy in mathematics education?* Paper presented at California State University at Northridge. <http://math.berkeley.edu/~wu/>.

# Chapter 4

## Using Curriculum-Based Measurement Fluency Data for Initial Screening Decisions

Erica S. Lembke, Abigail Carlisle and Apryl Poch

Curriculum-based measurement (CBM) has enjoyed a long history of success and study as a practice for data-based decision-making (Deno, 2003). Originally developed and studied at the University of Minnesota in the mid-1970s (see Shinn, 2012 or Tindal, 2013 for a detailed history), Stan Deno and his colleagues developed CBM measures and the problem-solving process as part of one of the Institutes for Research on Learning Disabilities (IRLDs), centers funded by the Office of Special Education Programs that addressed significant issues for students with learning disabilities. With Deno's interests in applied behavior analysis, it seemed logical to apply methodologies such as collecting baseline data, setting goals for students, and collecting and graphing ongoing data and then using them to make educational decisions, as a student's data is compared to a goal. As part of work in the IRLD, that is exactly what Deno and colleagues did, developing a system of technically adequate (i.e., reliable and valid) assessments that could be administered quickly and efficiently up to three times per week. These data would be graphed on an ongoing basis and compared with a goal set for a student. If data fell below the student's goal for a specified number of points, a curricular change or instructional tweak would be instituted. All of these components were couched in a problem-solving process so that teachers and teams could utilize on a frequent basis to help make better decisions about student learning. As you will note already, the CBM process

---

E. S. Lembke (✉) · A. Carlisle · A. Poch  
University of Missouri, Columbia, MO, USA  
e-mail: lembkee@missouri.ed

A. Carlisle  
e-mail: aaa961@mail.missouri.edu

A. Poch  
e-mail: Alpty9@mail.missouri.edu

© Springer Science+Business Media, LLC 2016  
K. D. Cummings, Y. Petscher (eds.), *The Fluency Construct*,  
DOI 10.1007/978-1-4939-2803-3\_4

or model is not just the measures themselves, but the use of those measures in a more comprehensive, problem-solving process. In this chapter the use of CBM, and specifically CBMs as measures of fluency, is discussed in depth. The theoretical support for measures of fluency is discussed along with more detailed research that supports the use of CBM, basic components of the process, and using CBM data to make screening decisions across a variety of academic subjects.

## Fluency as a Proxy for Academic Proficiency

Ask most educators to define what fluency is and many will say “fast reading or fast computing.” Defining fluency and providing a rationale for why fluency tasks might be important are critical, yet somewhat overlooked objectives. Fluency tasks are often associated with timing, working quickly, and are not always associated with a student’s best effort. Yet as this book illustrates, fluency is much more than just timed reading or math production. Fluency tasks embody characteristics of academic proficiency that students exhibit. Following administration directions, performance samples are elicited from students that are indicative of broader skills. For instance, some common CBM metrics include the number of words read correctly in a set amount of time or the number of mathematics problems completed during a given time. Both of these activities prompt students to work quickly (due to the timing) but also accurately; the final score represents the number correct, not just the number completed. When students have to work quickly and accurately, different skills are required than when they have unlimited time to complete a task. The cognitive skills accessed when students demonstrate fluent reading or computation are different from those accessed or applied on tasks where there is unlimited time or where accuracy is not paramount.

Fluency components in basic academic areas like reading, mathematics, and writing have been identified. In the area of reading, automaticity (LaBerge & Samuels, 1974), prosody (Schreiber, 1980), accuracy, and word recognition are all components that have been used in definitions of fluency (see Kuhn, Schwanenflugel, & Meisinger, 2010). In mathematics, the National Mathematics Panel (NMP, 2008) describes fluency in the area of whole numbers and fractions as critical foundational elements to prepare students for algebra. The NMP describes mathematical fluency as not just recalling basic facts, but being able to apply operational knowledge in problem solving. Fluency with algorithms for the basic mathematical areas is mentioned as well. In one paper, writing fluency is defined as the rate at which text is produced (Chenoweth & Hayes, 2001). Berninger and Fuller (1992) identified writing fluency as one of the key components in a two-part model to predict development skill and later writing achievement. Next, a brief discussion about the theory underlying each academic area is provided, prior to a return to how fluency tasks are utilized for CBM screening decisions.

## **Theoretical Support for Fluency as a Construct in Reading, Mathematics, and Writing**

### ***Reading***

In the area of reading, fluency measures have often been criticized because they appear to be simplistic “quick reads,” which only serve as an indication of how many words can be “called” in the time given. This lack of face validity has largely been overcome through careful discussion in the literature, as well as studies that address this issue head-on (c.f., Hamilton & Shinn, 2003).

Four main components have been utilized to describe fluent reading: automaticity, prosody, accuracy, and word recognition. Two major theories have emerged when linking word recognition with fluent reading: LaBerge and Samuels’ (1974) theory of automaticity and Perfetti’s (1985) verbal efficiency theory. Reading fluency is most frequently linked with the theory of automaticity, as described in LaBerge and Samuels’ (1974) seminal article. In this article, the authors described automaticity as rapid and fluent word reading. LaBerge and Samuels theorized that if children were able to read words more fluently, not much time would be spent decoding individual sounds and words. This, in turn, would free up working memory, leaving room for comprehension to take place. The more fluently a child reads the more working memory available and the better the comprehension. Similar to LaBerge and Samuels’ theory, Perfetti’s verbal efficiency theory proposes that readers can become more efficient readers through practice, and that efficiency in word recognition frees up cognitive resources. Perfetti posits that slow rates of word recognition “clog” working memory, affecting comprehension and recall. Shankweiler and Crain (1986) extended Perfetti’s verbal efficiency model by proposing that the combination of difficulties in orthographic decoding and limited-working memory capacity lead to difficulties in reading comprehension.

Moving from automaticity to prosody, Schreiber (1980) focuses on prosody, or expression, in reading and proposes that students’ lack of reading fluency may be a result of their inattention to prosodic cues, like phrasing and the rhythmic characteristics of language. Schwanenflugel, Hamilton, Wisenbaker, Kuhn and Stahl (2009) provide their definition of prosody, “when a child is reading prosodically, oral reading sounds much like speech with appropriate phrasing, pause structures, stress, rise and fall patterns, and general expressiveness” (p. 121). Meyer and Felton (1999) also include elements of prosody in their definition of reading fluency in a review of literature, where they describe fluency as “the ability to read connected text rapidly, smoothly, effortlessly, and automatically with little conscious attention to the mechanics of reading, such as decoding” (p. 284). Although Schreiber’s theory was proposed over 30 years ago, little research has been focused on prosody as an element of reading fluency. In a brief review for this chapter, out of 29 empirical articles on reading fluency interventions, only three included some form of prosody as a dependent measure. This search was conducted back to 2000 and involved a search of electronic databases using PsychInfo and Google Scholar but



did not involve a hand search. The majority of articles used dependent measures that addressed accuracy, fluency, and comprehension. There is a paucity of research using prosody as an indicator of fluency (Schwanenflugel, Hamilton, Wisenbaker, Kuhn, & Stahl, 2009) and this lack of use of prosodic features as an outcome is due in part to the difficulty of measuring expression in students' reading. For example, research by Young, Bowers, and MacKinnon (1996) and Young and Bowers (1995), in particular, have examined the effects of prosody on students' reading fluency using voice-activated devices to measure prosodic cues, such as pausal intrusions.

Another theory of reading fluency offered by Adams (1990) is a "connectionist" approach, in which orthography, phonology, meaning, and context interact to produce reading fluency. Adams suggests that rapid word identification and phrasal knowledge are necessary components, but are not sufficient to produce fluent reading on their own. Adams hypothesizes that a failure to make connections between words, meanings, and ideas results in nonfluent reading.

Wolf and Katzir-Cohen (2001) cite research by Kame'enui, Simmons, Good and Harn (2000) and Berninger, Abbott, Billingsley, and Nagy (2001) that characterizes fluency in a different manner than either a single-word recognition, prosodic, or connectionist view. Kame'enui and colleagues discuss fluency as a developmental process, where efforts at remediation need to be focused on early reading skills. Berninger and colleagues characterize fluency development as a systems approach, where the visual or verbal input, internal-language processes, and coordination of responses by the executive system all combine to influence growth in fluency.

The theory that underlies a researcher's position on fluency determines how studies are conducted and also what outcomes are measured. One of the keys to empirically examining a concept is operationalizing the term that you are studying. This has been difficult in the case of reading fluency, with definitions varying from study to study. Wolf and Katzir-Cohen (2001) discussed the lack of a clear definition of fluency and how even subtle changes in definition result in differences in assessment and intervention. Across all definitions, there are elements of speed and accuracy, including fluency described as verbally translating text with speed and accuracy (Fuchs, Fuchs, Hosp, & Jenkins, 2001) and accuracy of word recognition and reading speed, with an emphasis on speed (Samuels, 1997). Other researchers describe fluency as 3D, with expression or prosody accompanying rate and accuracy (Dowhower, 1991; Schreiber, 1980). Fluency has also been described as an indicator of comprehension.

When measuring or assessing fluency, nearly all studies use measures of reading speed and accuracy. This is not surprising given the theories just discussed and the emphasis on speed and accuracy as two of the primary components in addition to prosody. Reading speed is generally measured by counting the number of words that a student reads correctly in a constrained period of time, and accuracy is assessed by looking at the number of errors that a student makes in that reading. This is where CBM enters back into the picture.



## ***Curriculum-Based Measurement***

CBM is a system of progress monitoring in academic areas that utilizes technically adequate measures to assess progress. Technical adequacy studies are completed in three phases or stages (Fuchs, 2004). Stage 1, technical features of the static score (including reliability and validity), involves evaluation of measures that can be used to administer all students at limited times during the year to check on student performance compared to established norms. Stage 2, technical features of slope, involves development of measures that can be utilized for ongoing monitoring of student progress in an academic area. These progress-monitoring measures might be given as often as weekly. Stage 3, instructional utility, is focused on examining how the measures function when teachers utilize them for monitoring the progress of their students, including determining when instructional changes need to be made. Please see Burns, Silbergitt, Christ, Gibbons, and Coolong-Chaffin, Chap. 5, this volume, for more information about progress monitoring decisions—including response to intervention decisions.

CBM draws upon theories of automaticity and fluency, with a focus on development of measures that serve as indicators of broad constructs, such as reading proficiency. Deno, Mirkin, and Chiang (1982) initially identified the number of words read correctly in 1 min as a technically adequate indicator of overall reading proficiency. Initially, 1 min samples were collected to provide ease and efficiency for teacher administration. But there is nothing magic about 1-min timings; the duration of the assessment must be balanced with the item types and content of the measure. Thus, some CBM measures require 6 or 8 min to obtain adequate, reliable samples of information. More detail about average lengths of administration can be found in Table 4.1. In reading, repeated studies have been conducted on the efficacy of using the number of words read correctly in 1 min as a fluency indicator. It is important to note that the CBM research has shown the number of words read correctly in 1 min is not just a measure of decoding skill, but predicts general reading achievement. Multiple studies have been conducted on the validity and reliability of the CBM reading measure demonstrating strong correlations with other measures of fluency and comprehension (Shinn et al., 1992; Reschly et al., 2009).

## ***Mathematics***

In the area of mathematics, Rhymer et al. (2000) cites literature suggesting that computational fluency, defined as responding accurately and rapidly, leads to better long-term outcomes, maintenance of skills, and better application to novel mathematics tasks. The National Council of Teachers of Mathematics (NCTM, 2000) describes fluency in mathematics as "...having efficient and accurate methods for computing. Students exhibit computational fluency when they demonstrate *flexibility* in the computational methods they choose, *understand* and can explain these

**Table 4.1** Summary of common curriculum-based measurement (CBM) tools across the academic areas of reading, mathematics, writing, and other content domains

CBM probe	Description	Grade levels assessed	Time of administration	Procedure	Content or skill assessed	Scoring
<i>Reading</i>						
Letter naming fluency	Measures students' ability to rapidly name a selection of random lower and uppercase letters	Kindergarten	1 min	Individual	Naming orthographic letter symbols	Number of letters named correctly
Letter sound naming	Measures students' ability to accurately produce the phonological sound of the presented letter	Kindergarten	1 min	Individual	Naming most common sound for a given letter symbol	Number of correct letter sounds
Phoneme segmentation	Measures students' ability to accurately pronounce each phoneme of the presented word	Late Kindergarten	1 min	Individual	Segmentation of phonemes in words	Number of phonemes pronounced correctly
Nonsense word fluency	Measures students' ability to accurately segment or blend the sounds of a pseudo-word that primarily follows the CVC pattern	Grade 1	1 min	Individual	Letter-sound correspondence and phoneme blending	Number of sounds produced correctly
Word identification fluency	Measures students' ability to accurately read from a list of approximately 50 high-frequency words	Grade 1	1 min	Individual	Sight recognition of words	Number of words read correctly
Oral reading fluency (or passage reading fluency)	Measures students' ability to accurately and fluently read a brief passage at student's instructional level	Grades 1–8	1 min	Individual	Oral reading	Number of words read correctly per minute
Maze	Measures students' comprehension of a passage in which every seventh word is deleted; students must select the correct word from a series of distractors to fill in the missing word	Grades 1–6 Grades 7–12	1 min (elementary) 3 min (secondary)	Group	Comprehension	Number of words selected correctly

**Table 4.1** (continued)

CBM probe	Description	Grade levels assessed	Time of administration	Procedure	Content or skill assessed	Scoring
Vocabulary-matching	Measures students' ability to correctly match a set of content vocabulary; answer choices include two distractors	Middle school	5 min	Group	Vocabulary comprehension and knowledge	Correct vocabulary matches
<i>Mathematics</i>						
Oral counting	Measures students' ability to orally count out loud starting at one	Kindergarten and grade 1	1 min	Individual	Number counting	Numbers correctly counted
Number identification	Measures students' ability to rapidly name a series of randomly selected numbers between one and twenty	Kindergarten and grade 1	1 min	Individual	Number recognition	Numbers identified correctly
Quantity discrimination	Measures students' ability to accurately name the larger of two presented numbers	Kindergarten and grade 1	1 min	Individual	Discrimination of larger and smaller numbers	Number of correctly discriminated pairs
Missing number	Measures students' ability to accurately identify (name) the missing number in a sequence of three numbers; the missing number may be at the initial, medial, or final position	Kindergarten and grade 1	1 min	Individual	Recognizing a sequence and identifying the number needed to complete the sequence	Number of correctly identified missing numbers
Computation	Measures students' basic computation skills in single, mixed, or multi-step addition, subtraction, multiplication, and division	Grades 2–6	2–3 min	Group	Basic arithmetic (addition, subtraction, multiplication, division)	Number of correct digits
Concepts and application	Measures students' ability to correctly complete mathematical problems in an applied context	Grades 2–6	6–8 min	Group	Applied mathematics	Number of correct blanks

Table 4.1 (continued)

CBM probe	Description	Grade levels assessed	Time of administration	Procedure	Content or skill assessed	Scoring
Algebra	Four probes: basic skills—measures students' basic algebra performance; algebra foundations—measures students' knowledge of foundational algebra content; content analysis—measures students' knowledge of various algebraic concepts; translations—measures students' understanding of numerical relations across various formats	Middle school and high school	5–7 min	Group	Algebra	Basic skills—total number correct; algebra foundations—number of correct items; content analysis—sum of points across problems; translations—number of correct matches
<i>Writing</i>						
Word dictation	Measures students' transcription skills at the word level; ability to accurately spell words that have been dictated orally	Grades 1–6	2–3 min	Individual	Spelling	Words written, words spelled correctly, correct letter sequences, correct minus incorrect letter sequences
Picture word prompts	Measures students' transcription and text generation skills at the sentence level; ability to write sentences using pictures that are presented alongside the name of the picture	Grades 1–6	3 min	Group	Text generation at the sentence level	Words written, words spelled correctly, correct word sequences, correct minus incorrect word sequences
Story prompts	Measures students' transcription and text generation skills at the paragraph or discourse level; ability to write a story using a story prompt that reflects the experiences of students attending U.S. public schools	Grades 1–6 Grades 7–12	3 min (elementary) 5 or 7 min (secondary)	Group	Text generation at the passage level	Words written, words spelled correctly, correct word sequences, correct minus incorrect word sequences

methods, and produce accurate answers *efficiently*” (p. 152). In this definition, fluency is more than just production, but efficiency in application and explanation. Thomas (2012) suggests that perhaps we have moved beyond simple speed and accuracy in mathematics and that there are three competing definitions of fluency when applied to this skill area: “(1) Speed/efficiency are the sole components; (2) Speed/efficiency are the emphasized components, but meaning is also necessary; or (3) Meaning is the emphasized component and speed/efficiency are characterized as natural outgrowths of deep understanding” (p. 327). The National Mathematics Advisory Panel (2008) suggests that mathematical fluency includes both computational and procedural fluency. As many states implement the Common Core State Standards (CCSS), fluency is included as an aspect to be practiced, but not at the expense of understanding. The CCSS defines fluency as both quickness and accuracy. Within the standards, fluency is built from grade to grade on increasingly difficult skills. Clearly, there is a common theme throughout these reports, manuscripts, and standards indicating that rapid naming of facts and the ability to quickly apply procedures are critical to developing further mathematics skill.

In their article on computational fluency for high-school students, Calhoon et al. (2007) cite work demonstrating the far-reaching influences of fluency. For instance, The National Research Council (2001) provides an analogy suggesting that lack of computational fluency may have negative effects on mathematical comprehension similar to the effects that poor decoding has on reading comprehension (in Calhoon et al., 2007). In addition, Calhoon and her coauthors provide a overview of the literature suggesting that higher-order mathematics cannot be accessed as efficiently if fluency is not present (Gerber & Semmel, 1994; Johnson & Layng, 1994; Pellegrino & Goldman, 1987 in Calhoon, 2007). See also Clarke, Nelson, and Shanley (Chap. 3, this volume) for more information about the importance of fluency in mathematics assessments. The parallels between reading and mathematics fluency are compelling and provide a strong rationale for the use of fluency tasks as screening measures in systems of data-based decision-making. But more about that after we discuss theories of fluency in the area of writing.

## ***Writing***

In the area of writing, the skills most often targeted include transcription and text generation (McCutchen, 2006) or text production (Chenoweth & Hayes, 2001). Transcription, translation of language into written symbols, is most often measured through handwriting or spelling tasks for students. These tasks, which at first glance may appear to be fairly straightforward and perhaps even rudimentary, are strongly predictive of future writing performance (Graham, Harris, & Fink, 2000). In fact, in a recent study (Puranik & Alotaiba, 2012), the authors found that handwriting and spelling made statistically significant contributions to the prediction of written expression proficiency.

Text generation is the process of “turning ideas into words, sentences, and larger units of discourse” (McCutchen, 2006, p. 123). Text generation is also constrained by cognitive resources including working memory. How does working memory relate to writing fluency? Students who have strains on working memory may have difficulty retaining rules about use of grammar, accuracy in spelling, or simply brainstorming and retention of content ideas. Long-term memory resources are related to knowledge of topic and genre, which can constrain quality and quantity of text generation (McCutchen, 2006). Text generation has been found to be related to overall writing quality (Dellerman, Coirier, & Marchand, 1996) including that of beginning writers (Juel et al., 1986). Ritchey and colleagues (Chap. 2, this volume) have active research labs studying and refining theories supporting new and innovative measures of written language and are a great resource for further information on the topic.

Developing further understanding of the underlying theoretical constructs that support the use of fluency in each of these academic areas is important, as some would dismiss fluency tasks as simply “responding fast” if deeper understanding was not cultivated. As Deno and his colleagues at Minnesota, and others since, have reported, fluency with academic tasks serves as an indicator of something broader than just quick responding. As discussed in the preceding section, fluency with tasks can be directly linked to stronger comprehension in reading, greater problem solving ability in mathematics, and lengthier compositions in the area of writing. Theoretical foundations of learning support the use of fluency tasks for brief assessment in academic areas. With a better understanding of how fluency undergirds more sophisticated processing in these academic areas, we next move to a discussion regarding how fluency measures might be utilized for teachers and schools as part of a data-based decision-making, particularly in the area of universal screening.

## **Basic Components of Data-Based Decision-Making**

When CBM measures were developed in the mid-1970s, the initial framework for teacher-data utilization was termed as data-based program modification (DBPM; see Shinn, 2012 for a well-articulated account of this early work). DBPM had roots in teacher development, behavior analytic techniques, and precision teaching (Lindsley, 1990). Precision teaching included direct and explicit teaching methods, such as modeling and precise and frequent feedback using visual models. Deno and his colleague Phyllis Mirkin (1977) brought these components together in a manual that was published by the Council for Exceptional Children (CEC). The DBPM manual detailed methods that special education teachers could utilize to monitor the performance and progress of their students in basic skill areas. Special education teachers could empirically evaluate the progress of their students and make decisions about their instruction based on actual student performance data. This method differed from past practices where teachers might just guess about when to try something new or make judgments about the effectiveness of an intervention based on anecdotes or personal feelings. This new way of thinking brought about a

more data-based scientific approach to education. Centered around a problem-solving process (see Marston et al., 2003), DBPM provided a model to assist teachers as they identified an area of need, developed an intervention, monitored the progress of the student in the intervention, and then continued or modified the intervention after examination of data at regular intervals. This basic model is now termed data-based individualization (National Center on Intensive Intervention, 2012), Prevention Science (i.e., see Lembke, McMaster, & Stecker, 2009), or even response to intervention (RTI).

The basic components in the DBPM process include universal screening, goal setting, diagnostic assessment, hypothesis generation about potentially effective interventions, development and implementation of an instructional plan, weekly progress monitoring, and making ongoing changes in intervention using decision-making rules. Each step is described in more detail below using a case study of Mrs. Hammond's classroom and one of her students, Samuel.

### ***Step 1—Screening Using CBM Measures***

All students should be screened using CBM measures, ideally, three times per year (fall, winter, spring). Universal screening means that all students in a building are tested. Typical measures used for screening are short-duration tasks that are matched to students' grade levels; the results of those tests are then compared to established normative levels of performance. These norms are developed as a result of national, state, or local data collection, and translate into benchmark levels of performance that are standard criteria where students need to be performing to be deemed "not at risk" at a particular time of year (see Smolkowski, Cummings, and Stryker, Chap. 8 this volume, for more information about how benchmark levels of performance may be statistically determined). The criteria that determine risk status are determined statistically after examining data that has been collected for each grade at each time of year. Students who fall below a predetermined benchmark on the CBM are identified as needing additional instruction or interventions and their progress will be monitored more frequently.

#### **Case Study**

Mrs. Hammond teaches literacy to students in grades 1–3 who have difficulties or are on individualized education programs (IEPs) for reading or writing. At her school, these students come to work with her on their academic skills for at least 30 minutes per day in her classroom. Mrs. Hammond screens all of her students fall, winter, and spring using CBM measures. After scoring those measures, she has a sense of how her students compare to others who have completed these measures and she also has a better sense of the students' skill deficits. In the fall when she screens her students, she determines that several may have needs in the area of spelling.



### ***Step 2—Setting an Ambitious Goal for the Student and Labeling the Goal Line on the Student’s Progress Monitoring Graph***

An ambitious yet attainable goal is set for all of Mrs. Hammond’s students who scored below criterion. The specified goal is for a given time period (e.g., several months, a semester, end of year, etc.). Goals can be determined in one of three ways: (1) according to national norms, which vary by CBM product, (2) grade-level benchmarks, which also vary by CBM product or (3) an intraindividual framework, where a student’s individual data are used to project a reasonable goal in the time allotted using expected rates of growth. For example, using the intraindividual framework, a teacher could specify that a student would gain two words per week on a test of oral reading fluency (ORF). The end goal would be determined by the following formula:

$$\text{Goal} = (2 \text{ words per week expected gain}) \times 27 \text{ weeks left in the school year} \\ + \text{CBM screening score}$$

If a student had an initial score of 20, the goal score in 27 weeks would be 74.

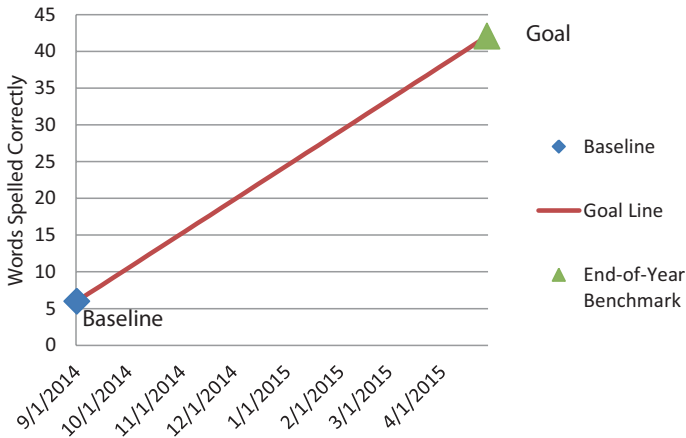
However the goal is decided, it along with the student’s current level of fluency (baseline; present level of performance) is marked on an individual student graph. The baseline and goal points are connected and a line is drawn between them. This goal line spans the number of instructional weeks between the baseline level of performance and the point by which the goal is desired to be achieved. This goal line determines the most direct route to take when attempting to reach the desired level of performance.

#### **Case Study**

Mrs. Hammond sets a goal for each student based on the national normative data available from the publisher of the measures and marks this goal on each student’s graph. She then marks each student’s baseline score, or current level of spelling fluency, and connects the two points to create a goal line (Fig. 4.1).

### ***Step 3—Identification of Strengths and Weaknesses Using Diagnostic Measures***

In addition to CBM screening measures, students who have been identified as requiring more intensive interventions may be given diagnostic measures. CBMs tell us *if* there is a problem. Diagnostic assessments tell us specifically *what* skills are deficient and what the student is able to do well. Diagnostic information is then used to develop an intervention plan and determine where to focus instruction. An example of a diagnostic fluency assessment is a miscue analysis (Fuchs, Fuchs, Hosp, & Jenkins, 2001), which determines the specific types of errors a student is making



**Fig. 4.1** Example progress monitoring graph with baseline level of performance, desired goal level performance (associated with a given time point in the school year), and the goal line

in reading. Teachers make notation about student errors as the student is reading aloud and then later go back and categorize the types of errors the student made.

### Case Study

Mrs. Hammond needs to decide *what* to teach for each student or group of students in her classroom who scored below the normative criterion at the screening point. She conducts an error analysis on each student's spelling to determine which letter patterns are in error. She groups students based on strengths and specific needs identified from the use of this diagnostic tool before initiating step 4.

### ***Step 4—Generating a Hypothesis About Appropriate Method to Individualize Instruction for the Student***

Using CBM results and diagnostic data as appropriate, educators should come up with logical ideas about what type of intervention program, instructional content, and delivery setting would be appropriate for each student. It is important to consider not only the specific skills the student needs to work on but also the amount and frequency of supplemental instruction and the size and composition of the intervention group.

### Case Study

Mrs. Hammond uses each student's school instructional plan or IEP as a template for how to individualize her lessons. She also brings back the error analysis data,

as well as any other information she has about the students, to bear in terms of selecting an intervention plan that is most likely to be successful. She considers that some students will likely do well in a small-group intervention setting, while a few students have significant needs that may be better served in a one-to-one or very small group structure. She also uses her own personal knowledge of student behavior to decide if students will work well together in their small groups. Perhaps Samuel and Sally tend to feed off each other in terms of who can be the biggest class clown; however, they work fine when paired with other students. She would then choose to separate them so that instructional time is more wisely used. After grouping considerations are finalized, the content of the lesson is made specific. We talk about this more in step 5.

### ***Step 5—Creating an Instructional Plan for Each Student or Group of Students***

Based on the above discussion, educators will develop an instructional plan with a goal and instructional activities for each student. These activities should be research- or evidence-based and typically include direct, explicit, and systematic intervention for the deficit area(s) identified during the diagnostic step above.

#### **Case Study**

To identify the content and activities she will use in the plan, Mrs. Hammond examines her menu of available intervention options matched to each student's needs and goals to further individualize for each student. In addition to determining the size of each student's intervention group, Mrs. Hammond thinks about whether a standardized intervention package that is delivered the same way to all students in a group might be appropriate, or if a more individualized strategy targeting specific spelling patterns may fit a student's needs. She must consider the intervention strategies and activities available, how much time each intervention will take, and the order of spelling patterns and targeted content she will cover with each student or group of students.

### ***Step 6—Beginning Regular and Frequent Instruction Using the Instructional Plan***

The instruction or intervention will be provided for as much time as possible, relevant to the skill needs. The greater the academic needs of the students, the more often the intervention should be implemented and for a greater length of time each session.

## Case Study

Mrs. Hammond begins implementing the instructional plan for each student. She monitors her own fidelity of implementation, keeping data on how long she is able to implement the plan each day and to what extent she is able to implement the essential elements of the plan. She also regularly and informally checks for mastery of content before moving on to the next unit or subunit in the instruction. This informal measurement is not CBM, but is critical to ensuring that instruction is effective for each student.

### ***Step 7—Regular Progress Monitoring, Including Scoring and Graphing, Using a CBM Measure***

To continuously monitor student response to the intervention, regular weekly progress monitoring data using a CBM is necessary. Continue to graph these data on the student's graph to determine if the student's performance is changing and how close his or her data points are to the goal line (see step 2 and Chap. 5 for more details about progress monitoring decision rules and evaluating a student's response to instruction). Weekly progress monitoring is recommended for students who are significantly behind their peers (e.g., someone in a tier 3-level intervention), whereas monitoring every other week or monthly may be more appropriate for students who are not as far behind (e.g., someone in a tier 2-level intervention).

## Case Study

In the area of progress monitoring, because she is providing an intensive tier 3 intervention for these particular students, Mrs. Hammond collects data weekly using one of the CBM measures in writing. She scores the measure and plots each individual data point for each student, by week. She then connects each of the progress monitoring data point for easier visual analysis. The connected line running through all of the student's data points is called a "trendline." This line describes the trend of student performance and can be interpreted relative to the aim or target line (shown in red, Fig. 4.2). This process of analyzing student performance over time is described in step 8.

### ***Step 8—Making Ongoing Changes in Instruction Based on Decision-Making Rules***

Using progress monitoring data, educators can determine if the instructional plan is having the intended effect. The main method for making educational decisions us-

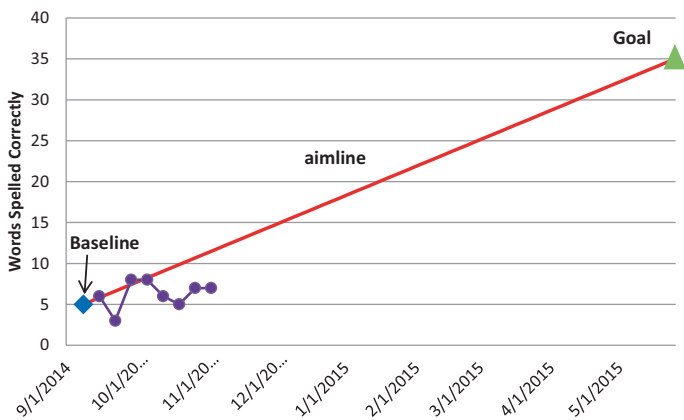


Fig. 4.2 Example progress monitoring with weekly student-level data points plotted and connected

ing student progress monitoring data is the trendline rule. There are several methods for determining the trend of student's progress. The National Center for Response to Intervention (NCII, 2014) lists the methods for several of the most common and supported in its glossary of terms available at the following web address: <http://www.intensiveintervention.org/ncii-glossary-terms>. Recent publications indicate that as many as 12 data points might be necessary to establish a reliable trend line (Ardoin et al., 2013). When the trendline is at or above the aimline, the intervention will be continued and likely faded if progress remains strong. When the student progress (represented by the trendline) is *below* the aimline, consider making a change to the intervention delivery or, in some cases, content.

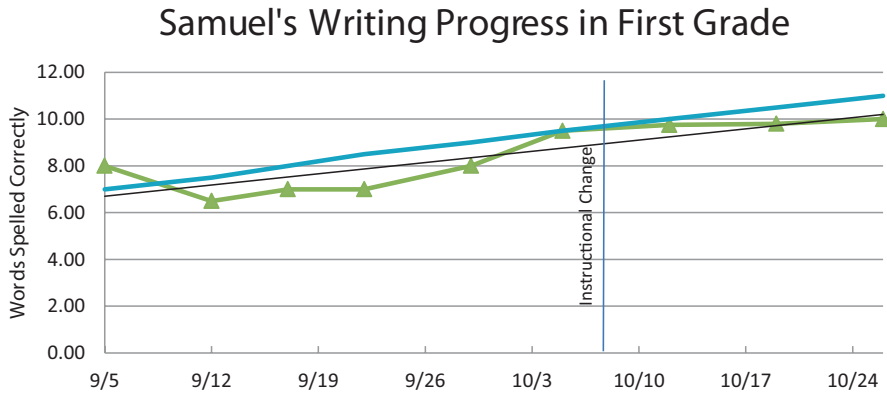
## Case Study

Mrs. Hammond uses decision-making rules to make ongoing decisions about intervention effectiveness. She does not want to continue with an intervention if student growth is not observed. For one of her students, Samuel, his data through October 1st indicates that his current trend of data is not approaching his aimline (see Fig. 4.3). For this reason, Mrs. Hammond decides to make a change for Samuel. She consults with her fellow teachers and her special education consultant to choose an intervention change that is supported by evidence.

For the purposes of this chapter, our focus remains on utilizing CBM measures for screening as specified in step 1.

## Using CBM for Screening Decisions

CBM measures embody specific characteristics, including: (a) efficient administration, (b) short duration, (c) technical adequacy, and (d) indicators of academic proficiency. The term *indicator* is used to signify the short duration of the measures



**Fig. 4.3** Progress monitoring graph for Samuel. *Solid line* indicates the aimline; the *line* specified by the data points (i.e., *triangles*) is Samuel's trendline of progress. The *vertical line* on the graph represents the point at which Mrs. Hammond's intervention change was made

as well as their strong relation to other measures of broad academic proficiency in that content area. Utilizing the theories underlying fluency, we can develop brief measures that serve as proxies for overall academic proficiency. Thus, although, a common measure of CBM in reading is the number of words read correctly in 1 min (oral reading fluency), this score serves as a broader indicator of academic proficiency in reading. As mentioned previously, in her 2004 article on the use of CBM measures, L. S. Fuchs described three stages of CBM research: stage 1, technical features of the static score; stage 2, technical features of slope; and stage 3, instructional utility. These stages are important because measures need to be researched and then utilized only for the purpose they were intended and the purposes for which they have been validated. A measure that is appropriate for stage 1 (assessing performance) may not have strong instructional utility. When considering what measure or combination of measures will be utilized for screening decisions for students, one must consider several technical features including: the accuracy of decision-making, predictive validity, and instructional utility of the measures across grades. In certain content areas like early mathematics (see Gersten et al., 2012), a battery of measures might be considered rather than a single measure.

### ***Other Features that May Impact Screening Decisions***

Once an appropriate measure is selected that maps on to our desired educational decision, other factors must be considered. The importance of classification accuracy is a critical component of any screener (Gersten, 2012; Johnson, Jenkins, & Petscher, 2010; Kovaleski, Vanderheyden, & Shapiro, 2013; Smolkowski et al., Chap. 8, this volume). Classification accuracy refers to how accurately a measure can be utilized to predict a decision regarding future student performance. For instance, classification accuracy might be calculated to determine how likely a student would be to pass or fail a high-stakes assessment in the spring based on initial performance on a

CBM during fall screening. Sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve are some of the statistics used to estimate the accuracy of a given screening measure, and is only interpretable for given associated set of cut points, in terms of correctly identifying students at one point in time as at at-risk or on track for outcomes measured at a later time. Sensitivity (i.e., the true positive fraction) describes how acutely a particular cut point on a screening measure identifies children as at-risk who end up failing the outcome measure; sensitivity is not interpretable without knowing the corresponding value of specificity for that same cut point. *Specificity* (i.e., 1-false positive fraction), on the other hand, refers to the degree to which a given cut point on specific screening measure rules out students who are not at risk for failing the outcome measure; a screener that is specific reduces the number of students who are identified erroneously as needing additional instructional support. There are trade-offs between sensitivity and specificity (i.e., as one increases the other decreases) and, depending on how accurate the screener is overall, the differences between sensitivity and specificity can be quite large. A book published in 2013 by Kovaleski and colleagues, *Identification of Students with Learning Disabilities Utilizing RTI*, provides a comprehensive overview of how classification accuracy can be improved and utilized for high-stakes decision-making and details how schools could use CBM screening data in a multitiered system to make classification decisions regarding students who might be in need of additional intervention. The goal for schools would be to maximize the number of students correctly identified. Utilizing a complementary process of screening and then a few weeks of follow-up progress monitoring to confirm or disconfirm the screening decision can be effective in enhancing classification accuracy.

This recent work in classification accuracy highlights the movement toward greater precision in decision-making utilizing CBM. Initial development of CBM focused on decisions that an individual teacher might make about a small group of students. A teacher would examine recent data values that had been collected and would apply a decision rule like the “three below, six above rule” (see Deno & Mirkin, 1982) where if three weekly data points were below the goal line (see Fig. 4.1), an intervention change would be needed, but if six weekly data points were *above* the goal line, it would be a time to raise the goal for the student. As the use of CBM morphed from individual teacher decision-making to decision-making for larger groups of students (utilizing normative data), the need for greater accuracy emerged. CBM essentially transitioned from serving as a key measure in an individualized, instructionally driven model for special education teachers, to being utilized across general education for universal screening, to serve as a key component in special education *eligibility* decision-making as part of an RTI model.

As another part of universal screening, CBM is utilized to predict performance on high-stakes tests (cf. McGlinchey & Hixson, 2004). Prediction of performance within a year or across years allows schools to better divert resources to students or groups of students who might fail if intervention is not provided. Thus, screening serves as an important technique to identify students at risk early, while there is still time to intervene. For CBM screening, the higher the stakes of the decision, the more important precision in decision-making becomes. For instance,



making a decision about student placement in special education is extremely high stakes and CBM screening data is one piece of data to aid that process. Precise decision-making is necessary when utilizing CBM data for this purpose, where a student's placement will be substantially influenced. A lower stakes decision that still requires specificity, but not to the same degree as special education evaluation, might be determination of small-group intervention activities for a low-performing classroom based on CBM screening data. The good news is that educators can find greater detail and more specificity on these issues in resources such as the book by Kovaleski et al. (2013) and in the chapters contained in this volume (i.e., Burns et al., Chap. 5; Smolkowski et al., Chap. 8; Espin & Deno, Chap. 13).

## Potential CBM Screening Measures

Next, discussion of the various measures available for screening will be provided. See Table 4.1 for a table of fluency-based CBMs in reading, math, writing, and other content areas that span the grade levels from early elementary through high school. The reader is also encouraged to access the Progress Monitoring and Screening Tools Charts assembled by the National Center on Intensive Intervention (NCII) and the National Center on RTI (NCRTI) respectively (see <http://www.intensiveintervention.org/chart/progress-monitoring> and <http://www.rti4success.org/resources/tools-charts/screening-tools-chart>) for additional information regarding the technical adequacy of the fluency-based CBMs discussed below. These charts are updated annually with new screening measures as well as with evolving information for existing tools.

### *Reading*

In the area of reading, typical measures utilized for screening in early literacy (i.e., pre-K through early grade 1) include letter naming, letter sound naming, phoneme segmentation, nonsense word fluency, and word identification fluency. Each measure is individually administered for 1 min and the number of correct responses is totaled and graphed. In the area of elementary literacy (i.e., grades 1–5), common CBM measures include oral reading fluency and maze. Oral reading fluency measures are individually administered for 1 min and the number of words read correctly in that minute is graphed. The maze task is group administered and the time for the task varies from 1 to 3 min depending on the students' grade level (e.g., earlier grade levels tend to have longer time to engage the task) and the specific test publisher. Students read a passage to themselves where every seventh word (approximately) is deleted and replaced by three choices: the correct word and two distractors. As students read the passage, they circle the word that they feel makes the most sense in the context of the passage. The number of correct choices selected

by the student is the score that is graphed. As students improve in their fluency on these tasks, an increase in academic achievement should also be noted.

Knowledge of letter names has been identified as one of the best predictors of later reading acquisition and growth (Stage, Sheppard, Davidson, & Browning, 2001). The letter naming fluency (LNF) CBM is a measure of students' ability to correctly name in 1 min a selection of random lowercase and uppercase letters. Each probe provides students with a randomly ordered list of letters and requires students to reproduce the letter name associated with each letter. The ability to accurately recognize and name different letters has been linked to later word reading ability as one must easily be able to recognize letters and sounds in order to use grapheme-phoneme knowledge to decode words (Stage et al., 2001).

CBMs in letter-sound naming (LSN) require students to produce the phonological sound of the presented letter. Students are given a list of 26 letters presented in random order and asked to say the sound that the letter makes. As the student reads off the list provided, the administrator marks errors on a corresponding score sheet (Fuchs & Fuchs, n.d.). Students receive one point for each correct letter sound.

Phoneme Segmentation CBM presents students with words containing approximately four phonemes. They must accurately pronounce each phoneme of the word presented. Students are timed for 1 min, whereas the administrator marks errors on a corresponding score sheet (Fuchs & Fuchs). Students receive one point for each phoneme pronounced correctly.

Nonsense word fluency is a first-grade dynamic indicators of basic literacy skills (DIBELS) measure (Good & Kaminski, 2002). Students are given 1 min to read through a list of pseudo-words that primarily follow the consonant-vowel-consonant (CVC) pattern. Credit is earned in one of two ways: (1) by saying each individual sound in the pseudo-word, or (2) by blending the sounds into a word (Fuchs, Fuchs, & Compton, 2004). Thus, the final NWF score is the number of sounds produced correctly, with up to three sounds per pseudo-word, as well as the total number of CVC words that were decoded completely and correctly. NWF therefore provides both an index of letter-sound correspondence and the ability to blend letters into words using the most common sounds for each letter (Fuchs, Fuchs, & Compton, 2004).

Word identification fluency (WIF) is a 1-min timed CBM for first-grade students that requires reading from a list of approximately 50 high-frequency words (Fuchs & Fuchs, [www.studentprogress.org](http://www.studentprogress.org)). On the scoring sheet, the administrator awards the student 1 point for reading a word completely and correctly and a 0 for an incorrect response, which could include an error on any part of the word. At the end of 1 min, the administrator circles the last word the student read, and tallies and then graphs the number of words read correctly. This score represents automatic word recognition, which is essential for reading proficiency (Fuchs, Fuchs, & Compton, 2004).

## ***Content Areas***

CBMs in the area of vocabulary matching for secondary students have also been researched as both indicators of performance and progress in social studies and

science (see Espin, Shin, & Busch, 2005; Espin, Busch, Shin, & Kruschwitz, 2001; special issue of *Assessment for Effective Intervention* on content area measurement 38, Lembke, E.). These studies have examined both student-read and administrator-read forms. Student-read forms contain 22 vocabulary terms with two distractors printed on the left side of a page and listed alphabetically. Twenty definitions were provided on the right side of the same page; each definition was reworked to contain 15 words or fewer. Vocabulary terms were chosen at random from a social studies textbook and from teachers' lectures. Vocabulary-matching probes were administered for 5 min. Students were expected to read both the vocabulary terms and the definitions and to match each term with its respective definition. Administrator-read probes were developed from the same list of vocabulary words, but the form students received only contained the vocabulary terms. Administrators read the definitions aloud, one at a time, to students, who were asked to identify the term that best matched the definition read. Probes were administered for 5 min with a 15-s interval between each item. Students received one point for each vocabulary term matched correctly. Espin, Shin, and Busch (2005) found that student-read probes produced reliable and valid growth trajectories and exhibited sufficient sensitivity to growth over time.

## ***Mathematics***

In mathematics, the most common measures at the elementary level have traditionally been computation or concepts and applications measures (see Foegen et al., 2007 for a detailed review). These measures require students to complete simple arithmetic or applied mathematics problems. In early numeracy development (i.e., kindergarten through grade 1), CBMs in oral counting, number identification, quantity discrimination, and missing number are commonly used. These measures capture early numeracy skills that are believed to be related to later mathematical proficiency and understanding and are based on the principle of number sense (Clarke et al., Chap. 3, this volume; Clarke & Shinn, 2004). CBMs for secondary math instruction in the area of Algebra have also been developed and are discussed at the end of this subsection.

The oral counting CBM requires students to count out loud starting at one and going as far as they can in 1 min. No student materials are required; the administrator records the student's progress on a scoring sheet, placing a bracket after the last number that the student states. The final score is determined as the number of correct values, in a sequence, that the student was able to say. This value is recorded and graphed. Only numbers counted in sequence are counted as correct. Numbers not counted in sequence, and numbers provided to the student after a brief hesitation (e.g., 3 s) are scored as incorrect (Clarke & Shinn, 2004).

Number identification is another early numeracy CBM that requires students to verbally identify, or name, numbers between 1 and 20 that are presented in random order. Students are provided with a form that contains a table of numbers and are asked to read the numbers aloud, reading from left to right across rows. Numbers

correctly identified are scored as correct; numbers misidentified or numbers that are skipped are marked as incorrect. Students pausing for 3 or more seconds are prompted by the administrator to move onto the next number. The number of correctly identified numbers is recorded and graphed (Clarke & Shinn, 2004).

Quantity discrimination asks students to, when presented with two numbers, verbally state which is larger. Numbers are randomly paired and appear side by side in separate boxes. Students are asked to work from left to right across rows identifying the larger number. Boxes in which the student correctly identified the larger number are scored as correct. When students select the smaller number, state an incorrect answer, or hesitate for more than 3 s, an error is marked. As with other CBMs, when the student hesitates for at least 3 s, he or she is prompted by the administrator to move onto the next pair. The number of correctly discriminated pairs is totaled and then recorded (Clarke & Shinn, 2004).

Missing number measures ask students to identify a missing number within a sequence of three, with the missing number appearing at either the initial, medial, or final position. The three-string sequences are presented in individual boxes, and students complete the task with paper and pencil. Students need to correctly identify the missing number in the sequence to receive credit for the response. Responses are scored as incorrect if the student either names the incorrect number or skips a problem. Students who hesitate for at least 3 s are directed by the administrator to move onto the next sequence (Clarke & Shinn, 2004).

Computation CBM assess students' basic computation skills in single, mixed, or multi-step addition, subtraction, multiplication, and division (Lembke & Stecker, 2007). This CBM is group administered for 2–3 min (Fuchs, Fuchs, & Zumeta, 2008). Students receive credit for correctly identified digits when completing each problem; thus, partial credit is possible for more advanced items with two or more digits in the final answer.

Concepts and application measures assess students' skills with completing mathematical problems in an applied context. Included domains in these measures vary by grade level, but can include counting, number concepts, number naming, measurement, money, grid reading, charts, graphs, fractions, decimals, applied computation, word problems, quantity discrimination, temperature, etc. (Fuchs, Fuchs, & Zumeta, 2008; Lembke & Stecker, 2007). Often math concepts and application measures include multiple digits or words in their complete answers. Students receive credit for the number of blanks completed correctly, allowing them to earn partial credit for their responses.

CBM probes in Algebra, part of Project AAIMS (see [http://www.education.ia-state.edu/c\\_i/aaims/](http://www.education.ia-state.edu/c_i/aaims/)), have been identified and include four different probes. The basic skills algebra probes contain 60 items and are designed to test a student's basic algebra performance in areas including, but not limited to, solving simple equations, combining like terms, applying the distributive property, working with integers, and working with proportions. (Johnson, Gallow, & Allenger, 2013; Foegen & Morrison, 2010). The basic skills probes are group administered and students have 5 min to complete as many items as they can. Students earn credit (1 point) for each correctly answered problem (60 points maximum); the total number correct are then tallied and graphed (Foegen & Morrison, 2010; Foegen, Olson, & Impecoven-Lind, 2008).

Beyond basic skills, the algebra foundations CBM is group administered for 5 min and assesses student performance across the following domains: (a) variables and expressions, (b) integers, (c) exponents, (d) order of operations, (e) graphing, (f) solving simple equations, (g) extending patterns in data tables, (h) writing a word phrase for expressions, and (i) graphing expressions. Students earn credit (1 point) for each correct item (50 points maximum; Foegen & Morrison, 2010; Foegen, Olson, & Impecoven-Lind, 2008).

The third AAIMS measure, content analysis, is a multiple-choice CBM that covers numerous algebraic concepts (e.g., solving equations, evaluating expressions, solving linear systems, calculating slope, simplifying expressions with exponents; Foegen & Morrison, 2010; Foegen, Olson, & Impecoven-Lind, 2008). Each problem is worth a total of 3 points. Students earn full credit by circling the correct choice. Partial credit is awarded for showing work using a rubric-based key. Scores are the total sum of points across all problems. Students are provided 7 min to complete as many items as they can (Foegen, Olson, & Impecoven-Lind, 2008).

The final algebra measure is translations. This task requires students to explore numerical relations in multiple formats (e.g., data tables, graphs, equations; Foegen, Olson, & Impecoven-Lind, 2008). Students are required to correctly identify matches across the multiple formats. The algebra measures are currently under further development through federal grant work conducted by Foegen and colleagues (see [http://www.education.iastate.edu/c\\_i/aaims/](http://www.education.iastate.edu/c_i/aaims/)).

## **Writing**

CBM in writing originally involved story prompts to which students responded for 3–5 min and were scored for number of words written (WW), words spelled correctly (WSC), and correct word sequences (CWS, which accounts for spelling and grammar; Videen, Deno, & Marston, 1982). These measures have yielded reliable and valid indices of writing proficiency for students in grades 2 and up (see McMaster & Espin, 2007 for a review). Recently, researchers have extended writing CBMs to provide indicators of students' early writing proficiency with evidence of reliability, validity, and sensitivity to growth made over short-time periods (e.g., Coker & Ritchey, 2010; Lembke, Deno, & Hall, 2003; Hampton, Lembke, & Summers, 2010; McMaster, Du, & Petursdottir, 2009; McMaster, Du, Yeo, Deno, Parker, & Ellis, 2011; Parker, McMaster, Medhanie, & Silberglitt, 2011).

CBM for beginning writers has included tasks designed to capture transcription and text generation to reflect early writing development at the word, sentence, and discourse levels of language and has included scaffolding (in the form of verbal, picture, or written prompts) to support young writers' developing self-regulatory skills. The tasks are timed to gauge production fluency, which is a strong predictor of overall writing quality (e.g., Berninger & Swanson, 1994) most likely because fluency in lower-order processes frees up cognitive resources for higher-order processes (Berninger & Amtmann, 2003; McCutchen, 2006). These tasks have included dictation, sentence writing, and story writing. The writing subsection of Table 4.2

provides a summary of research on CBM for beginning writers, highlighting measures that have been established as having adequate reliability, validity, and utility for monitoring progress over time across early elementary grades. Three CBM tasks that are well established in terms of reliability, validity, and capacity to show growth in short-time periods are word dictation, picture-word prompts, and story prompts. These measures offer teachers a selection of tools that can be utilized at the word, sentence, and discourse levels based on the grade and skill level of their students.

Word dictation (WD), a measure designed to capture students' transcription skills at the word level, is administered individually. WD requires students to write words dictated by the examiner. Words (approximately 20–40) used in these probes may come from high-frequency word lists designed to address students' knowledge of various spelling patterns (e.g., VC, CVC, VCe, etc.), grade-level spelling texts, or unit-specific words.

Picture word (PW) prompts are group administered and are designed to capture students' transcription and text generation skills at the sentence level. Each prompt contains a series of pictures with the corresponding name below the picture. Students are asked to compose a sentence about the picture and the names of each picture may be read aloud to students prior to administration.

Story prompts (SP), also group administered, are designed to capture students' transcription and text generation skills at the paragraph or discourse level. Each prompt contains a story starter surrounding a topic that reflects the experiences of students attending the US public schools. They contain simple vocabulary and a simple sentence structure. Students are presented with the story prompt and asked to think about their story for 30 s before responding. Elementary aged students are then asked to write independently for 3 min. Secondary level students may write for 5 or 7 min, but the time of administration must remain constant across the academic year.

## **Issues for Consideration, Including Limitations of the Use of CBM for Screening Decisions**

### ***Face Validity and Fluency***

Educators should use CBMs strategically, realizing that these measures are important for quick screening but other measures may need to be brought to bear in cases where students are identified as needing additional support. In this way, multiple skills in a content area are assessed in order to gain a more robust picture of student ability and draw reasonable conclusions about his or her overall performance. For example, assessing oral reading fluency or letter identification provides one piece of information about a student's reading ability, but additional measures of reading comprehension as well as diagnostic measures to investigate types of errors made are necessary to make sound instructional decisions for students. Educators should use CBMs with an eye toward interpreting results carefully and within the confines of what the task requires students to do.



## ***Students with Speech and Language Impairments***

In general, when administering CBMs that require a verbal response to students with speech and language impairments, educators must score and interpret assessment results with caution. Students should not be penalized for errors of production if those errors are a direct result of their speech impairment. Likewise, educators should be careful when interpreting CBM results for students with language impairments, as they may be slower to process directions and give responses. CBM results can provide information about a student's performance, but should not be the only piece of data used to make educational decisions regarding classroom performance for students with speech and language impairments. In addition, consideration should be given to the individual needs of students and whether fluency-based assessments accurately capture students' performance and progress, keeping in mind that in some cases, a fluency-based measure is not appropriate.

For students with speech difficulties, the person administering the fluency CBM should be familiar with the student's speech patterns and be able to correctly score his/her responses. It may be necessary for the school or district's speech-language pathologist to participate in testing or to serve as a second scorer. Additionally, students with a fluency impairment (i.e., a student who stutters) may be accurate in his/her oral reading fluency but may read slowly. These students may have a low rate of words read per minute (WPM) as a result of their dysfluency, not as a result of a true reading deficit. It is important to consider this when making educational decisions and grouping students by ability, as an oral reading fluency CBM may not be the best representation of the reading level of a student who has difficulties or disabilities with respect to speech production.

Students with language impairments may read fluently but may in fact struggle with comprehension and vocabulary of a CBM passage. Educators must be sure to assess the comprehension of students with language impairments and use data from reading fluency and comprehension measures to determine the need for reading interventions.

In all content areas, including mathematics and writing, students with language impairments may struggle to understand and correctly follow the directions for a CBM task, especially the first time the assessment is given. Every effort must be taken to ensure that CBM results reflect the student's ability in that content area and not the deficits created by their language impairment. Although administration directions are standardized to allow for comparison of results across peers, it may be necessary to repeat or even reword the task directions, depending on the age of the student and the severity of the language impairment. If directions were altered in any way, educators must interpret results carefully and avoid making peer comparisons (e.g., avoid comparing scores collected in that manner to established criterion-referenced goals or benchmarks). Rather, in cases where the standardization of the assessment is lost, only within-individual comparisons can be made (e.g., comparing a student's current performance to her past performance given consistent breaks in standardization between the two administrations). Educators should use several different, and sometimes individualized, assessments to make educational and intervention decisions for students with language impairments.



## **Future Research**

Using CBM fluency data for universal screening can provide a snapshot of a student's academic proficiency in reading, mathematics, writing, and other content areas. Although these tasks have demonstrated adequate reliability and validity and provide a general indication of a student's academic health as it relates to broader academic skills in each of these areas, future research surrounding the use of fluency measures is needed in many areas. Consistent with previous sections of this chapter, issues related to future research are also broken down by skill and content areas.

### ***Reading***

Although research in CBM in reading is well documented, many areas remain open for future research. Specifically, research might address ways to map reading rates to productive reading strategies, text type (e.g., narrative vs. expository), level of text difficulty for secondary students, as well as the kinds of qualitative data that can be extracted from fluency measures “to help teachers generate diagnostically useful performance profiles,” including linking diagnostic information to instructional recommendations, and exploring methods for assessing prosody and its impact on reading competence (Fuchs, Fuchs, Hosp, & Jenkins, 2001, p. 252).

### ***Mathematics***

In early mathematics, we need more research regarding whether a single mathematical indicator can be utilized across multiple grade levels to track student progress and growth over time *or* whether multiple measures of early mathematical fluency are required for assessing progress (Clarke & Shinn, 2004). At the secondary level, more research is needed surrounding the technical adequacy of CBM in advanced mathematics, such as algebra or geometry, along with the “instructional utility” (Calhoun, 2008, p. 237) of these measures for teachers making instructional decisions. Future research might also explore the criterion validity of M-CBM and high-stakes assessments, as well as the criterion and predictive validity of multiple-skill M-CBMs (measures with several types of mathematics tasks on one probe) (Christ & Vining, 2006).

### ***Writing***

What defines fluency and how to define fluency in writing continues to remain an area for future research. Though quantitative scoring indices have primarily been

utilized throughout the early research in this area (Coker & Ritchey, 2010) in which researchers have demonstrated that simple, countable indices of writing such as WW, WSC, and CWS are reliable and valid (Deno, Marston, Mirkin, Lowry, Sindelar, & Jenkins, 1982; Deno, Mirkin, & Marston, 1982; Videen, Deno, & Marston, 1982), the validity of these writing indices have not remained stable across grade level. Furthermore, relatively little work explores alternative scoring indices in writing or the use of qualitative writing indices. Whether such scoring (both quantitative and qualitative) is consistent with the indices that teachers value most must also be considered (Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002).

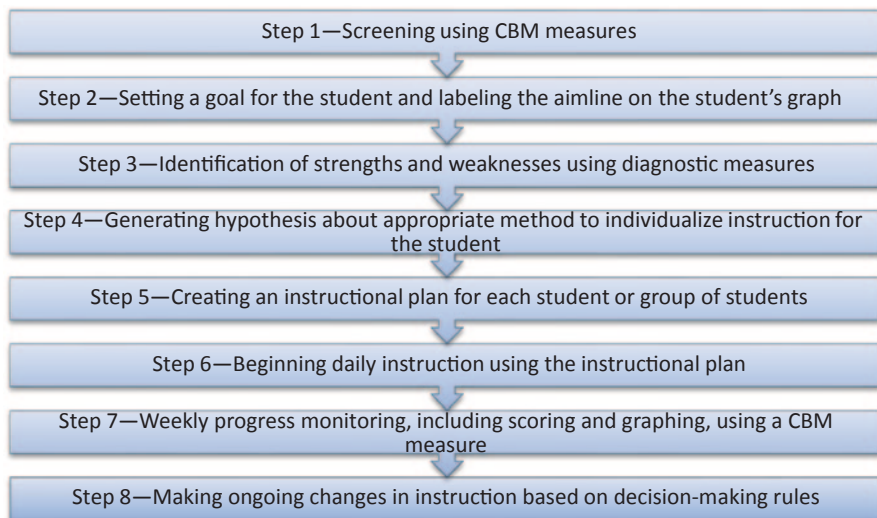
### *Content Areas*

In the content areas of social studies and science, CBM vocabulary assessments offer promise, however, relatively little work exists in this area. Although research has begun to address how vocabulary assessments might be used for making placement decisions, identifying discrepancies in student performance and progress, and determining need for intervention (Espin, Busch, Shin, & Kruschwitz, 2001; Espin, Shin, & Busch, 2005), additional empirical research is needed. Research is also needed in using vocabulary-matching measures as progress-monitoring measures, on how CBM in vocabulary influences teacher decision-making, and on student achievement in the content areas (Espin, Busch, Shin, & Kruschwitz, 2001; Espin, Shin, & Busch, 2005).

Moreover, fluency tasks in the above outlined areas share many similar challenges that necessitate future research. Although research must continue across all of Fuchs' (2004) stages given the varying depths of the extant literature in the different areas, specific attention is needed in stage 3. Namely, as Espin, Shin, and Busch (2005) noted above, fluency CBMs must explore the effect of both screening and progress monitoring on teachers' instructional practices and student performance (see also Foegen & Morrison, 2010, regarding the effect of teacher use of CBM on student progress), as well as on special education decision-making. As Fuchs (2004) adds, incorporating teacher and student feedback loops for designing "instructionally informative diagnostic profiles" as supplements to graphing, may improve "CBMs instructional utility" (p. 191). If the goal is that use of CBM will result in improved outcomes for students, research must ensure that teachers know how to use and interpret CBM scores to inform their instruction. Similarly, teachers must perceive the data obtained through the use of CBM as useful, what Calhoon<sup>1</sup> (2008) calls the "acceptability or utility" (p. 237) of CBM. Acceptability can be particularly difficult at the middle and high-school levels where teachers' caseloads often exceed 100–180 students (Calhoon, 2008). The unique needs of secondary teachers and students must be examined, as certain CBM measures and scoring techniques

---

<sup>1</sup> Although Calhoon talks specifically of the struggles in mathematics, these concerns must be recognized across content classes at the secondary level.



**Fig. 4.4** Steps in a data-based model for decision-making using CBM

may not be appropriate for adolescent learners, failing to adequately capture their individual progress.

Furthermore, the movement toward multiple-skill CBM measures over single-skill measures, has been supported, as adequate growth over time in the latter may not be sufficient for demonstrating broader knowledge in the content domain (Fuchs, 2004). Fuchs (2004) also suggests that long-term progress monitoring using single-skill measures may too narrowly restrict teachers’ instructional focus.

Finally, while Tindal and Parker (1991) recommend “embed[ding] assessment within a decision-making framework” (p. 218) for writing, such a recommendation is also important for the other fluency measures. Situating and supporting instructional, intervention, and placement recommendations within and with data-based decisions is central to identifying what to measure, how to improve performance, and how to document student growth to ensure that the decisions being made accurately reflects student need (Tindal & Parker, 1991; See Fig. 4.4 for specific steps).

As fluency is a complex construct, research must continue to explore the many nuances of what it means to be fluent in reading, mathematics, writing, and the content areas across grade levels and across ability levels of students. Unfortunately, such measures will only be useful to the extent that they are properly used, interpreted, and provide valuable instructional recommendations for teachers. Though the current research demonstrates great promise, much work remains.

## References

- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge: MIT Press.
- Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of recommendations and research surrounding Curriculum Based Measurement of Oral Reading Fluency (CBM-R) decision rules. *Journal of School Psychology*. doi:10.1016/j.jsp.2012.09.04.
- Berninger, V. W., & Amtmann, D. (2003). Preventing written expression disabilities through early and continuing assessment and intervention for handwriting and/or spelling problems: Research into practice. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 345–363). New York: Guilford.
- Berninger, V. W., & Fuller, F. (1992). Gender differences in orthographic, verbal, and compositional fluency: Implications for assessing writing disabilities in primary grade children. *Journal of School Psychology*, 30(4), 363–382.
- Berninger, V. W., & Swanson, H. L. (1994). Modifying Hayes and Flower's model of skilled writing to explain beginning and developing writing. In J. S. Carlson (Series Ed.) & E. C. Butterfield (Vol. Ed.), *Advances in cognition and educational practice, Vol. 2: Children's writing: Toward a process theory of the development of skilled writing* (pp. 57–81). Greenwich: JAI Press.
- Berninger, V. W., Abbott, R. D., Billingsley, F. & Nagy, W. (2001). Processes underlying timing and fluency of reading: Efficiency, automaticity, coordination, and morphological awareness. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 383–414). Timonium: York Press.
- Calhoon, M. B. (2008). Curriculum-based measurement for mathematics at the high school level. *Assessment for Effective Intervention*, 33(4), 234–239.
- Calhoon, M. B., Emerson, R. W., Flores, M., & Houchins, D. E. (2007). Computational fluency performance profile of high school students with mathematics disabilities. *Remedial and Special Education*, 28(5), 292–303.
- Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, 18, 80–98.
- Christ, T. J., & Vining, O. (2006). Curriculum-based measurement procedures to develop multiple-skill mathematics computation probes: Evaluation of random and stratified stimulus-set arrangements. *School Psychology Review*, 35(3), 387–400.
- Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Today*, 33(2), 234–248.
- Coker, D. L., & Ritchey, K. D. (2010). Curriculum-based measurement of writing in kindergarten and first grade: An investigation of production and qualitative scores. *Exceptional Children*, 76(2), 175–193.
- Dellerman, P., Coirier, P., Marchand, E. (1996). Planning and expertise in argumentative composition. In G. Rijlaarsdam, H. van den Bergh, & M. Couzijn (Eds.), *Theories, models, and methodology in writing research* (pp. 182–195). Amsterdam: Amsterdam University Press.
- Deno S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Reston: Council for Exceptional Children.
- Deno, S., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. (1982). *The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study* (Vol. IRLD-RR-87). Minnesota: University of Minnesota, Institute for Research on Learning Disabilities.
- Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, 49(1), 36–45.
- Deno, S. L., Mirkin, P., & Marston, D. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children Special Education and Pediatrics: A New Relationship*, 48, 368–371.

- Dowhower, S. L. (1991). Speaking of prosody: Fluency's unattended bedfellow. *Theory into Practice, 30*(3), 165–175.
- Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*, 184–192.
- Espin, C. A., Busch, T. W., Shin, J., & Kruschwitz, R. (2001). Curriculum-based measurement in the content areas: Validity of vocabulary-matching as an indicator of performance in social studies. *Learning Disabilities Research and Practice, 16*(3), 142–151.
- Espin, C. A., Shin, J., & Busch, T. W. (2005). Curriculum-based measurement in the content areas: Vocabulary matching as an indicator of progress in social studies learning. *Journal of Learning Disabilities, 38*(4), 353–363.
- Foegen, A., & Morrison, C. (2010). Putting algebra progress monitoring into practice: Insights from the field. *Intervention in School and Clinic, 46*(2), 95–103.
- Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *Journal of Special Education, 41*(2), 121–139.
- Foegen, A., Olson, J. R., Impeccoven-Lind, L. (2008). Developing progress monitoring measures for secondary mathematics: An illustration in algebra. *Assessment for Effective Intervention, 33*(4), 240–249.
- Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188–192.
- Fuchs, L. S., & Fuchs, D. (n.d.). Using CBM for progress monitoring. National Center of Student Progress Monitoring. <http://www.studentprogress.org/>. Accessed 08 April 2014.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239–256.
- Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71*(1), 7–21.
- Fuchs, L. S., Fuchs, D., & Zumeta, R. O. (2008). A curricular-sampling approach to progress monitoring: Mathematics concepts and applications. *Assessment for Effective Intervention, 33*(4), 225–233.
- Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measures in writing. *School Psychology Review, 31*(4), 477–497.
- Gerber, M. M., & Semmel, D. S. (1994). Computer-based dynamic assessment of multidigit multiplication. *Exceptional Children, 61*, 114–126.
- Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children, 78*(4), 423–445.
- Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills*, 6th edn. Eugene: Institute for the Development of Educational Achievement.
- Goodman, Y. M., Watson, D., & Burke, C. (1987). *The reading miscue inventory*. Katonah: Richard C. Owen.
- Graham, S., Harris, K. R., & Fink, B. (2000). Is handwriting causally related to learning to write? Treatment of handwriting problems in beginning writers. *Journal of Educational Psychology, 92*(4), 620. doi:10.1037/0022-0663.92.4.620.
- Hamilton, C. R., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review, 32*, 228–240.
- Hampton, Lembke, & Summers. (2010). *Examining the technical adequacy of early writing curriculum-based progress monitoring measures* (Unpublished manuscript). Columbia: University of Missouri.
- Johnson, K. R., & Layng, T. V. (1994). The Morningside model of generative instruction. In R. Gardner III, D. M. Sainato, J. O. Cooper, T. E. Heron, W. L. Heward, J. Eshleman et al. (Eds.), *Behavior analysis in education: Focus on measurably superior instruction* (pp. 173–197). Monterey: Brooks/Cole.

- Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention, 35*(3), 131–140. doi:10.1177/1534508409348375.
- Johnson, E. S., Galow, P. A., & Allenger, R. (2013). Application of algebra curriculum-based measurements for decision making in middle and high school. *Assessment for Effective Intervention, 39*(1), 3–11.
- Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology, 78*(4), 243.
- Kame'enui, E. J., Simmons, D. C., Good, R. H., & Harn, B. A. (2000). The use of fluency-based measures in early identification and evaluation of intervention efficacy in schools. In M. Wolf (Ed.), *Time, fluency, and dyslexia* (pp. 307–333). Parkton: York Press.
- Kovaleski, J. F., Van Der Heyden, A. M., & Shapiro, E. S. (2013). *The RTI approach to evaluating learning disabilities*. New York: Guilford.
- Kuhn, M., Schwanenflugel, P., & Meisinger, E. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*, 230–251. <http://dx.doi.org/10.1598/rrq.45.2.4>. Accessed 15 Apr 2014.
- La Berge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293–323.
- Lembke, E. & Stecker, P. (2007). *Curriculum-based measurement in mathematics: An evidence-based formative assessment procedure*. Portsmouth: RMC Research Corporation, Center on Instruction. <http://files.eric.ed.gov/fulltext/ED521574.pdf>. Accessed 09 Apr 2014.
- Lembke, E., Deno, S. L., & Hall, K. (2003). Identifying an indicator of growth in early writing proficiency for elementary school students. *Assessment for Effective Intervention, 28*(3–4), 23–35. doi:10.1177/073724770302800304.
- Lembke, E.S., McMaster, K., & Stecker, P.M. (2009). The prevention science of reading research within a response-to-intervention model. *Psychology in the Schools, 47*(1), 22–35.
- Lindsley, O. R. (1990). Precision teaching: By teachers for children. *Teaching Exceptional Children, 22*(3), 10–15.
- Marston, D., Muyskens, P., Lau, M., & Canter, A. (2003). Problem-solving model for decision making with high-incidence disabilities: The Minneapolis experience. *Learning Disabilities Research and Practice, 18*(3), 187–200.
- McCutchen, D. (2006). Cognitive factors in the development of children's writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 115–130). New York: Guilford.
- McGlinchey, M. & Hixson, M. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*, 193–203.
- McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *Journal of Special Education, 41*(2), 68–84.
- McMaster, K. L., Du, X., & Petursdottir, A. (2009). Technical features of curriculum-based measures for beginning writers. *Journal of Learning Disabilities, 42*, 41–60.
- McMaster, K. L., Du, X., Yeo, S., Deno, S. L., Parker, D., & Ellis, T. (2011). Curriculum-based measures of beginning writing: Technical features of the slope. *Exceptional Children, 77*(2), 185–206.
- Meyer, M. S., & Felton, R. H. (1999). Repeated reading to enhance fluency: Old approaches and new directions. *Annals of Dyslexia, 49*(1), 283–306.
- National Center on Intensive Intervention (NCII, 2012/2014). *Office of Special Education Programs, National Center on Intensive Intervention*. U.S. Department of Education: Washington, DC.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston: Author.
- National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: US Department of Education.
- National Research Council. (2001). Looking at mathematics and learning. In J. Kilpatrick, J. Swafford, & B. Findell (Eds.), *Adding it up: Helping children learn mathematics* (pp. 1–16). Washington, DC: National Academy Press.



- Parker, D. C., McMaster, K. L., Medhanie, A., & Silbergitt, B. (2011). Modeling early writing growth with curriculum-based measures. *School Psychology Quarterly*, 26(4), 290–304. doi:10.1037/a0026833.
- Pellegrino, J. W., & Goldman, S. R. (1987). Information processing and elementary mathematics. *Journal of Learning Disabilities*, 20(1), 23–34.
- Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.
- Puranik, C. S., & Alotaiba, S. (2012). Examining the contribution of handwriting and spelling to written expression in kindergarten children. *Reading and Writing*, 25(7), 1523–1546.
- Reschly, A., Busch, T., Betts, J., Deno, S., & Long, J. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology*, 47(6), 427–469.
- Ritchey, K. D., & Speece, D. L. (2006). From letter names to word reading: The nascent role of sublexical fluency. *Contemporary Educational Psychology*, 31, 301–327.
- Rhymer, K. N., Dittmer, K. I., Skinner, C. H., & Jackson, B. (2000). Effectiveness of a multi-component treatment for improving mathematics fluency. *School Psychology Quarterly*, 15(1), 40.
- Samuels, S. J. (1997). The method of repeated readings. *Psychology*, 6, 293–323.
- Schreiber, P. A. (1980). On the acquisition of reading fluency. *Journal of Literacy Research*, 12(3), 177–186.
- Schwanenflugel, Hamilton, Wisenbaker, Kuhn, & Stahl. (2009)
- Shankweiler, D., & Crain, S. (1986). Language mechanisms and reading disorder: A modular approach. *Cognition*, 24(1), 139–168.
- Shinn, M. R. (2012). Reflections on the influence of CBM on educational practice and policy and its progenitor. In C. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.), *Measure of success: The influence of curriculum-based measurement on education* (pp. 341–356). Minneapolis: University of Minnesota Press.
- Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement: A confirmatory analysis of its relationship to reading. *School Psychology Review*, 21, 459–479.
- Stage, S. A., Sheppard, J., Davidson, M. M., & Browning, M. M. (2001). Prediction of first-graders' growth in oral reading fluency using kindergarten letter fluency. *Journal of School Psychology*, 39(3), 225–237.
- Thomas, J. N. (2012). Toward meaning-driven mathematical fluency. *School Science and Mathematics*, 112(6), 327–329.
- Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education (International Scholarly Research Network)*. doi:10.1155/2013/958530.
- Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research and Practice*, 6(4), 211–218.
- Videen, J., Deno, S. L., & Marston, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression*, Vol. IRLD-RR-84. Minnesota: University of Minnesota, Institute for Research on Learning Disabilities.
- Young, A., & Greig Bowers, P. (1995). Individual difference and text difficulty determinants of reading fluency and expressiveness. *Journal of Experimental Child Psychology*, 60(3), 428–454.
- Young, A. R., Bowers, P. G., & MacKinnon, G. E. (1996). Effects of prosodic modeling and repeated reading on poor readers' fluency and comprehension. *Applied Psycholinguistics*, 17(1), 59–84.
- Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading*, 5(3), 211–239.



## Chapter 5

# Using Oral Reading Fluency to Evaluate Response to Intervention and to Identify Students not Making Sufficient Progress

**Matthew K. Burns, Benjamin Silbergitt, Theodore J. Christ,  
Kimberly A. Gibbons and Melissa Coolong-Chaffin**

Reading is complex to learn, teach, accurately practice, and understand (Carnine, Silbert, Kame'enui, & Tarver 2009). Thus, instructional reading methods have long been venomously argued and debated in schools, articles, books, newspapers, government panels, and policy papers with few resounding agreements (Coles, 1998). The complexity of the reading process and the importance of effective reading instruction serve to emphasize the need for accurate reading assessments in order to develop instructional techniques that maximize every student's individual skills and ultimately allow them to become effective readers.

An effective reading assessment should measure some of the basic foundational components of reading development and instruction. The National Reading Panel (NRP, 2000) report responded to a Congressional mandate to help educators and policymakers identify the central skills necessary for reading acquisition. After critically reviewing more than 100,000 peer-reviewed and scientifically based studies, the NRP identified the now famous five building blocks for reading instruction of phonemic awareness, phonics, fluency, vocabulary, and text comprehension. Reading fluency is the ability to read with both speed and accuracy (NRP, 2000) and

---

M. K. Burns (✉)  
University of Missouri, Columbia, MO, USA  
e-mail: burnsmk@missouri.edu

B. Silbergitt  
Technology and Information Educational Services, Falcon Heights, MN, USA  
e-mail: Benjamin.Silbergitt@ties.k12.mn.us

T. J. Christ  
University of Minnesota, Minneapolis, MN, USA  
e-mail: tchrist@umd.edu

K. A. Gibbons  
St. Croix River Education District, Rush City, MN, USA  
e-mail: kgibbons@scred.k12.mn.us

M. Coolong-Chaffin  
University of Wisconsin Eau Claire, Eau Claire, WI, USA  
e-mail: chaffimc@uwec.edu

fluent readers read effortlessly without having to devote cognitive resources toward decoding words, which allows the reader to focus on word recognition, comprehension of text ideas, and the use of their own background knowledge to connect with concepts within the text. Reading fluency is an important skill that allows the reader to both gain meaning from the text and subsequently reinforce reading due to the success of the process. Thus, fluency is a skill that most directly encompasses the other reading skills within it which makes it an effective assessment approach (Fuchs, Fuchs, Hops, & Jenkins, 2001).

## **Reading Fluency as an Assessment**

Fluent reading is often used as a general outcome measure for overall reading health because it is an indication of proficient decoding and comprehension (Berninger, Abbott, Vermeulen, & Fulton, 2006; Fuchs, Fuchs, Hops, & Jenkins, 2001). Curriculum-based measurement (CBM) of oral reading fluency (ORF) consists of a set of standardized, individually administered reading probes that are brief (1 min) to administer and measure reading development across multiple grades. ORF assesses a child's reading skills by measuring the accuracy and fluency of connected text with grade-level material, and constraining the measurement procedure to 1 min standardizes the process and makes the data comparable across students and points in time while also facilitating efficient measurement of reading.

### ***Purposes of Assessing ORF***

The ORF assessments consist of multiple standardized sets of passages and administration procedures that are designed to identify children who may require additional instructional support and to monitor progress towards predetermined instructional goals. ORF works well as a screener because it is theoretically linked to overall reading competence (Fuchs, Fuchs, Hosp, & Jenkins, 2001) and it correlates well with reading comprehension ( $r = .48-.55$ , Valencia et al., 2010;  $r = .54$ , Burns et al., 2011;  $r = .76$ , Roberts, Good & Corcoran, 2005). Moreover, ORF data predict state accountability test performance in reading with high accuracy (Shapiro, Keller, Lutz, Santoro, & Hintze, 2006; Yeo, 2009).

### ***Psychometric Properties of ORF***

Various studies have suggested that ORF assessments from CBM have acceptable technical adequacy. A synthesis of psychometric evaluations of ORF data collected with CBM found test-retest, interrater, and alternateform reliability coefficients that

ranged from .82 to .95 (Goffreda & DiPerna, 2010). The criterion-related validity of ORF data is also well established, with correlation coefficients that exceed .60 between it and other measures of reading (e.g., comprehension, vocabulary, decoding, and word identification) (Reschly, Busch, Betts, Deno, & Long, 2009). Moreover, estimates of predictive validity also exceeded .60, even when the duration between predictor and criterion was 1 year (Keller-Margulis, Shapiro, & Hintze, 2008).

## ORF Assessment and Response to Intervention

Response to intervention (RtI) has made assessment and data-based decision-making important priorities in K-12 schools. Federal law states that a local educational agency “may use a process that determines if the child responds to scientific, research-based intervention as a part of the evaluation procedures” (Pub. L. No. 108–446 § 614 [b][6][A]; § 614 [b][2 & 3]), which is the legal basis for RtI data to be used to identify a specific learning disability (SLD). Yet, RtI generally refers to a systematic process of collecting data to assess and evaluate the effectiveness of instruction and intervention, and not necessarily to make special education eligibility decisions. Our preferred definition is that RtI is *the systematic use of assessment data to efficiently allocate resources to enhance learning for all students* (Burns & VanDerHeyden, 2006).

Given that RtI is a process to make decisions in order to allocate resources, there are four potential decisions with RtI data. Riley-Tillman and Burns (2009) described these decisions. First, an effective intervention is identified and successful. In other words, the intervention led to increased growth and the student demonstrated adequate proficiency. In this case, the intervention is often no longer needed (decision 1). Second, the intervention is effective but not successful. In this case, the student’s reading skills increases at a desirable rate, but the student still remains below expectations for proficiency and the intervention often continues (decision 2). The third possibility is a twist in the second because it involves an intervention that is effective, but is so intense that it cannot continue without increased resources (decision 3). Finally, the fourth possibility is that an effective intervention is not identified (decision 4).

The first two options involve continued or decreased resources, but the latter two involved increased resources including possibly, but not necessarily, special education. Thus, RtI is more of a resource-allocation model than an approach for identifying special education eligibility. In most RtI models, special education is prescribed when research-based interventions are attempted for a predetermined amount of time without demonstrated success, but few states have well-established criteria for determining the effectiveness of an intervention.

Effectiveness of an intervention within an RtI framework is typically evaluated in terms of the students’ level and rate of growth on ORF assessments with CBM (Gresham, 2002). As stated above, ORF data are reliable and correlate well with other important measures of reading (Goffreda & DiPerna, 2010;

Reschly et al., 2009; Yeo, 2009), but that research was conducted with only single ORF data points. The reliability and validity of rate of growth metrics for ORF should also be considered when evaluating intervention effectiveness. Below we discuss two approaches to interpreting ORF data when monitoring the rate of growth of student progress.

## **Interpreting ORF Data when Monitoring Student Progress**

Historically, ORF data were presented in a time-series graph and a collection of the student's individual progress data points were compared to an aimline with data-point decision rules used to guide decisions about whether or not the gains were sufficient (Deno, 1986). However, recent reforms due to RtI implementation have led to student progress being evaluated by computing and interpreting a numerical slope rather than using a simple data-point decision rule as a part of a dual discrepancy that evaluates both the value of the slope of growth *and* the post-intervention reading level (DD; Fuchs, 2003).

### ***Data-Point Decision Rule***

The aimline, sometimes referred to as a goal line (Deno, 1986), is the expected rate of progress of a student given where the student started and the desired post-intervention goal. More specifically, an aimline is the line that connects the initial level (baseline) of performance and the desired level by a particular goal date (typically the middle or end of a school year). Student data are then plotted on a time-series graph and progress is measured by comparing subsequent individual or groups of data points to the aimline. A student is making sufficient progress if the data points approximate the aimline, with specific decisions about the adequacy of their gains evaluated using data-point decision rules. The most common data-point decision rule uses a requirement of three consecutive data points above the aimline to suggest that a more ambitious goal is needed and three or four consecutive data points below the aimline to suggest that the intervention is not effective (Mirkin, Deno, Tindal, & Kuehnle, 1982; Fuchs, Fuchs, Hintze, & Lembke, 2006; Shinn, 1989). Within an RtI framework, three consecutive points below the aimline suggest that the intervention is not working and a more intensive corresponding tier of intervention is needed (i.e., decision 4; Fuchs, Fuchs, Hintze, & Lembke, 2006).

The process of student self-monitoring with plotted CBM data has been demonstrated to improve reading achievement (Stecker & Fuchs, 2000) and teachers' use of progress monitoring data using data-point decision rules has resulted in both more frequent revisions to student education plans and increased student achievement (Fuchs, Fuchs, Hamlett, & Stecker, 1991; Stecker, Fuchs, & Fuchs, 2005). Still, most data-point decision rules are based on expert opinion rather than em-

pirical research (Ardoin, Christ, Morena, Cormier, & Klingbeil, 2013) and RtI has enhanced the need for research regarding decisions made from progress monitoring data because those data could result in changes to student placement or resource allocation (Salvia, Ysseldyke, & Bolt, 2010).

Using data-point decision rules for resource allocation could also be problematic because the resulting decisions tend to be unreliable and fall far below standards for data used to make resource allocation decisions. Previous research found a split-half reliability estimate for a data-point decision rule that recommended three data points below the aimline of .44, which is well below desired standards (Burns, Scholin, Kosciulek, & Livingston, 2010). Three potential reasons why comparing individual progress monitoring data points to aimlines might have lesser reliability are that: (a) similar growth rates for different students could result in different decisions based on the level of baseline performance (VanDerHeyden, Witt, & Barnett, 2005), (b) the standard error of measurement for the single data point from which the aimline is drawn can be too large to make aimlines useful (Burns et al., 2010), and (c) graphed data do not result in reliable estimates of growth until 12 to 14 data points are collected (Christ, 2006). Still, most test authors and practitioner recommendations continue to focus on data-point decisions being made with only three data points while ignoring all other data.

### ***Dual Discrepancy***

ORF data used to monitor progress can also be examined from a dual discrepancy (DD) model (Burns & Senesac, 2005; Fuchs & Fuchs, 1998; Vellutino et al., 1996). Growth within a DD framework is conceptualized as a linear function represented by least-squares regression slopes between ORF progress measures and days of instruction (Deno, Fuchs, Marston, & Shinn, 2001). Thus, growth data can be computed with ordinary least squares (OLS). Students whose performance is discrepant from established standards in both post-intervention level *and* growth (as measured by slope, not a cluster of data points) are described as dually discrepant and the intervention is judged to be ineffective. Students whose post-intervention fluency scores meet or exceed standards are identified as proficient readers who are no longer in need of intervention (i.e., decision 1). Growth data suggesting that the student is making sufficient progress, but has yet to meet ORF standards, such as a seasonal benchmark, suggest that the intervention is leading to sufficient growth and should be continued (i.e., decision 2).

### **Dual Discrepancy Criteria**

Although there are a number of published studies that provide guidance to define the criteria for a DD model within RtI, there are serious questions left to be answered before a standardized three-tiered model with DD criteria can be established

(Burns & Ysseldyke, 2005). Achievement discrepancies are defined in a variety of ways within the published research. A discrepancy in *level* has been defined as performance below the 25th percentile (standard score of 90) on post-intervention, norm-referenced reading tests (Torgesen et al., 2001). Discrepancies in level can also be defined by performance below criterion-referenced standards on post-intervention ORF assessment administrations (Fuchs, 2003) such as estimates of risk (e.g., the goal criteria used by the Dynamic Indicators of Basic Early Literacy Skills (DIBELS); Good, Gruba & Kaminski, 2002).

Similar methods are used to define discrepancies in the rate of academic growth. Published standards for growth are within the range of 1–2 words read correctly per minute (WRCM) per week (Deno, Fuchs, Marston, & Shin, 2001; Fuchs, Fuchs, & Speece, 2002; Hasbrouck & Tindal, 2006) and are based on local normative data to derive percentile cutoffs (Fuchs, Fuchs, McMaster, & Al Otaiba, 2003), group medians (Vellutino et al., 1996), or rates of growth associated with effective practice (e.g., 1.39 WRCM per week; Deno, Fuchs, Marston & Shin, 2001). Normative slope standards have been suggested such as the 25th percentile (Burns & Senesac, 2005) or one standard deviation (Fuchs, 2003; Speece & Case, 2001; Speece, Case, & Molloy, 2003) of the growth rates for the general population. Students in higher grades have higher ORF scores and lower rates of typical growth than students in lower grades (Hasbrouck & Tindal, 2006). For example, a first grade student whose score increased at a rate of 1.2 WRCM per week would score at the 50th percentile for his or her grade group, but a score increase of only 0.90 WRCM per week would represent the 50th percentile for a fifth grade student. Normative expectations for growth with ORF data generally expect students to improve by 1–2 WRCM (Deno, Fuchs, Marston, & Shin, 2001; Silberglitt & Hintze, 2007).

There is clearly no consensus for how to best determine adequate rates of growth within a DD model with various pros and cons for each. A criterion-referenced approach to both level and slope has also been explored. Previous research used slope standards identified by Deno and colleagues (2001) as the criterion for slope of student learning (i.e., Burns, 2007; Burns & Senesac, 2005). This approach to DD used criterion-referenced standards for both post-intervention reading level (DIBELS standards) and slope of student learning (1.39 WRCM per week as identified by Deno, Fuchs, Marston, & Shin, 2001).

A hybrid approach can be used in place of solely a criterion-referenced or normative approach to DD. Interpretive criteria could include (a) a normative criterion for both level and slope, (b) a criterion-referenced approach for level and slope, (c) a norm-referenced approach for level and a criterion-referenced approach for slope, or (d) a criterion-referenced approach for level and a normative approach for slope. Silberglitt and Gibbons (2005) proposed a norm-referenced approach where discrepancies in level were defined as CBM-R performances below the 7th percentile on local norms. Although Silberglitt and Gibbons implemented a norm-referenced approach to evaluate level, they maintained a criterion-referenced approach to evaluate the rate of growth. But it is often not practical or feasible to conduct progress monitoring with all students to obtain reliable slopes and to then use normative slope data requires cross-cohort comparisons. To address these concerns, Silberglitt and Gibbons (2005) used target scores for creating a slope criterion.

First, benchmark target scores were developed and linked to performance on state-mandated assessments using logistic regression (Silbergliitt & Hintze, 2005), which resulted in consistent benchmark standards across a school year because all target scores were linked to the same outcome variable. Next, the rate of growth of these target scores was examined. In the above example, 34 weeks elapsed between fall and spring benchmark testing, so the weekly growth rate of the targets was 1.03 WRCM per week. This growth rate represents a criterion equal to 1 year's growth in 1 year's time (Silbergliitt & Gibbons, 2005). Finally, a confidence interval around this criterion was computed and students below this confidence interval were considered significantly below criterion (to assist in eligibility decisions), while students above the confidence interval were considered significantly above criterion. Students within the confidence interval required additional data in order to make a decision about their status. For DD, children whose post-intervention level fell below the 7th percentile (normative for level) and whose slope of learning fell below the confidence interval around the criterion for their grade level (criterion-referenced for slope) were considered not responding to their current instructional program (Silbergliitt & Gibbons, 2005).

## Evaluating Criteria to Interpret ORF Data within a Dual Discrepancy

Previous research found that a DD approach is superior to a single-discrepancy approach, which relies on either a low reading level *or* rate of achievement (Fuchs, 2003). The results of some studies provided evidence that DD criteria for identification and diagnosis converged with the outcomes of norm-referenced reading tests because students who exhibited a DD scored lower on reading tests than students with reading difficulties who did not exhibit a DD (Burns & Senesac, 2005; McMaster, Fuchs, Fuchs, & Compton, 2005; Speece & Case, 2001; Speece, Case, & Molloy 2003). In fact, several DD models consistently differentiated between students who are responding and students who are not responding adequately to their current instructional environment (Burns & Senesac, 2005).

There are a variety of options for establishing DD criteria within an RtI model. Given that high stakes decisions will be made with these data, it seems especially important to refine and validate the models used. Thus, we examined four DD models including each of the following combinations to determine nonresponse: (a) criterion-referenced level and norm-referenced slope at the 25th percentile; (b) criterion-referenced level and a norm-referenced slope at the 16th percentile (i.e., 1 *SD* below the mean); (c) norm-referenced level at the 7th percentile and criterion-referenced slopes; (d) and, finally, criterion-referenced level and criterion-referenced slopes. Our goal was to determine: (a) for children identified as at-risk for reading failure, do the four DD models differentiate reading skills of those who are dually discrepant and those who are not, and (b) what would the prevalence rate be for children identified as dually discrepant within the four DD models?



## *Comparing the Data*

Students ( $n = 3354$ ) from second through eighth grades in five rural school districts in Minnesota were participants for the study and all data were gathered during a single school year. There were 1625 (48.4 %) females and 1729 (51.6 % males), with 3156 (94.1 %) of the students being Caucasian and 32.3 % participating in the federal free or reduced-price lunch program. Native American students made up 2.4 % of the sample, and Asian-American (1.1 %), African-American (1.4 %), and Hispanic (1.2 %) children each represented less than 2 % of the sample. Moreover, there was a relatively equal distribution of students across grade levels with the exception of Grade 8 (2nd = 14.1 %, 3rd = 14.8 %, 4th = 16.6 %, 5th = 17.4 %, 6th = 16.8 %, 7th = 12.8 %, and 8th = 7.6 %).

The sample was limited to those students whose spring benchmark ORF score fell at or below the 25th percentile ( $n = 773$ ). The sample was limited to the bottom 25 % because that level of student performance was sufficiently discrepant in level to identify a child as at-risk for reading difficulties.

## *Measures and Criteria*

### **ORF**

Post-intervention reading levels were assessed with the spring benchmark assessment using CBM of ORF. Student slopes were calculated using the three ORF measures that are administered in the fall, winter, and spring of the academic year. At each of the three assessment periods, students were asked to read three ORF standardized reading passages (AIMSweb, 2006) that were written at a grade-appropriate difficulty level and were standardized and equated as recommended by Howe and Shinn (2002). Individual passage WRCM scores consisted of the total words read correctly in 1 min minus the total number of errors made while reading (see Wayman, Wallace, Wiley, Tichá, & Espin, 2007 for more information about general scoring rules for ORF). Final data for this study were recorded as WRCM and the median score of the three assessments was used for decisions and analyses.

Trained school personnel administered and scored the ORF probes using standardized procedures (Shinn, 1989). The training consisted of a 2-h instructional session followed by a competency assessment that involved a scoring rehearsal. Each scorer was required to come within two correct words per minute of the correct score on three consecutive videotaped assessments. The size of the sample did not allow for inter-rater reliability data to be collected, but previous research on ORF data collected with CBM demonstrated inter-rater reliability above .99 and test-retest reliability that exceeded .90 (Goffreda & DiPerna, 2010).

## Measures of Academic Progress for Reading

As consistent with previous research (Burns & Senesac, 2005), a standardized measure that was external to the curriculum and the DD criteria was used to compare reading skills of students identified as dually discrepant and those who were not. The Measures of Academic Progress for Reading (MAPR; Northwest Evaluation Association; NWEA, 2003) was administered to every student in the study three times per year and was used to compare the reading skills as an external criterion to evaluate the decisions. The MAPR is a criterion-referenced assessment designed with item response theory to determine item difficulty and equate different items from a pool of several thousand (NWEA, 2003). Normative data were also provided for students in Grades 2 to 10, based on the national pool of test takers (NWEA, 2005). The test primarily consisted of brief passages, each of which was followed by a multiple-choice question, with 42 items on the test. The MAPR is untimed, but typically requires approximately 1 h to complete.

After completing the test, students are assigned a Rasch Unit (RIT) Score. RIT scores are based on student performance and difficulty level of the items, and range from a low of approximately 130 to a high of approximately 270. Thus, MAPR data are RIT scores that are grade-level independent (NWEA, 2005). In other words a fifth grade student and third grade student with equal scores demonstrate equal reading skills. Moreover, since data were grouped across grades, grade-specific standard scores were calculated (mean = 100, standard deviation = 15) and then used for all analyses. Test-retest reliability estimates for the MAPR data ranged from .81 (second grade) to .89 (fourth grade), estimates of internal consistency all exceeded .90, and correlation coefficients between MAPR and other group reading tests all exceeded .75 (NWEA, 2004).

## Dual Discrepancy Criteria

Four DD models were used for the study. The first two used the same criterion level and different normative slopes. The criterion level was the grade-appropriate standard from DIBELS, in that children who scored at or below the cutoff score for at-risk were considered to be below the benchmark level. The normative slope consisted of slopes that fell below the 25th percentile for the grade (henceforth referred to as the 25th percentile model) or that fell more than one standard deviation below the grade-level mean (henceforth, referred to as the 1SD model).

The third DD model used a normative level (below the 7th percentile) and criterion-referenced slope (Silbergitt & Gibbons, 2005). The criterion slope was computed by determining the progress necessary to achieve benchmark target scores (Silbergitt & Gibbons, 2005). Thus, children were identified as dually discrepant if their post intervention ORF score was at or below the 7th percentile and the slope fell below the lowest end of the confidence interval around the target slope. Subsequently, this will be referred to as the Silbergitt and Gibbons model.

Finally, the fourth DD model used a criterion-referenced level (the DIBELS standards) and criterion-referenced slope (1.39 WRC/min per week for effective practice; Deno, Fuchs, Marston, & Shin, 2001). Scores represented a DD if the ORF level fell below the grade appropriate cutoff for at-risk in DIBELS standards and slopes fell below 1.39 WRC/min per week (Deno, Fuchs, Marston, & Shin, 2001). Subsequently, this will be referred to as the Deno et al. model.

## *Procedure*

ORF data were collected for all students as part of the systematic benchmark reading assessments collected three times each year in the participating districts (see Howe, Scierka, Gibbons, & Silberglitt, 2003 for a complete discussion of the assessment plans). The MAPR was administered by trained classroom teachers or by other trained school personnel using standardized procedures (NWEA, 2003).

The slope of student growth was computed across the fall, winter, and spring benchmark ORF assessments using ordinary least squares regression. There were minimal missing data (less than 5 %), which were assumed to be missing at random. Individual slope estimates for each student were calculated using SAS version 9 (2005). The assessments were conducted in September, January, and May, and resulting slope estimates represented the average weekly increase in WRCM. The equation used to estimate performance for each individual  $i$  at time  $j$  was as follows:

$$y_{ij} = b_{0i} + b_{1i}t_{ij} + e_{ij} \quad (5.1)$$

From this equation, the value of  $\beta_{1i}$  for each individual  $i$  served as the datum for that student. The subsequent analysis was designed to evaluate trend. Consistent with DD models for service delivery, students who are discrepant in both level and trend are considered for more intensive services.

Data for the 773 students were then used to group children as responding or not responding using the four different DD models. MAPR standard scores between the two groups were compared with an analysis of variance and a Bonferroni-corrected alpha level of .0125. Because the scores for two groups were compared, a Cohen's  $d$  (1989) was computed to examine the magnitude of the difference, which was interpreted with Cohen's criteria of 0.80 being large, 0.50 medium, and 0.20 small.

The second research question inquired about prevalence of children identified as DD. Therefore, the number of children who were demarked as DD within each of the four models was recorded and divided by the number of children in the entire sample, which was the entire second through eighth grade population for the five districts.

**Table 5.1** Mean standard scores for Measures of Academic Progress for Reading between dually discrepant and non-dually discrepant students for the four dual discrepancy criteria

Criteria	Dual discrepant			Non-dual discrepant			<i>F</i>	<i>d</i>
	<i>n</i>	Mean	SD	<i>n</i>	Mean	SD		
25th percentile	181	76.63	15.47	589	87.21	13.44	79.75*	0.73
1 SD	104	74.30	16.50	669	86.37	13.61	66.63*	0.80
Deno et al.	379	79.19	15.29	387	90.10	11.71	123.12*	0.81
Silberglitt & Gibbons	76	70.17	15.10	697	86.34	13.66	93.98*	1.12

Mean standard score for norm group was 100 (SD = 15)

\* $p < .001$

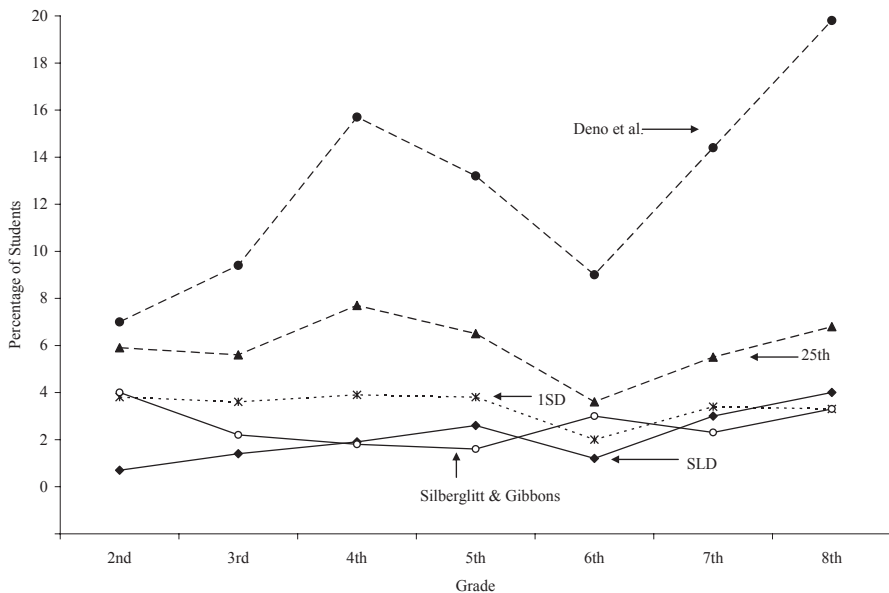
## Results

Our first goal was to inquire if the four DD models differentiated reading scores between students who were DD and those who were not, using a standardized norm-referenced measure of reading (MAPR). As displayed in Table 5.1, all four models significantly differentiated between the two groups. The Deno et al. model, and the Silberglitt and Gibbons model both led to large effects (0.81 and 1.12, respectively). The 1SD and 25th percentile models both used a criterion-referenced level and norm-referenced slope, so the *d* for the two were averaged and equaled a medium to large effect of 0.77.

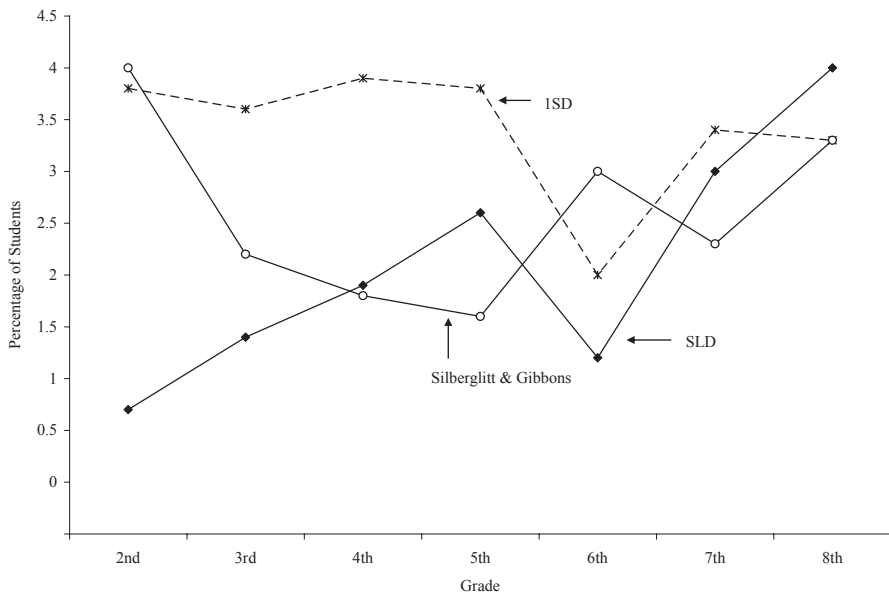
The second goal addressed prevalence of children identified as DD using the four models. As shown in Fig. 5.1, the overall prevalence rate of students who were DD ranged from less than 2 to 7 % in second grade, and from approximately 3 to 20 % in eighth grade. The range in prevalence was 3.6–7.7 % for the 25th percentile model, 2–3.9 % for the 1SD model, 1.6–4 % for Silberglitt & Gibbons model, and 7–20 % for the Deno et al. model. A total of 77 children (2.2 %) were identified as DD by all four models. The range for the prevalence of children identified as SLD using a traditional IQ-achievement discrepancy model per existing district guidelines was 0.7–4 %.

The two DD models with the smallest range were the 1SD model and the Silberglitt & Gibbons model. To better visually examine prevalence rates, these two models were graphed separately in Fig. 5.2 along with the SLD prevalence for comparison. As shown in the figure, the percentage of children identified as DD was approximately 4 % in second grade for both models, and fell to approximately 3.5 % in eighth grade. The percentage of children identified as SLD was 0.7 % in second grade, but increased to 4 % in eighth grade. Therefore, the prevalence rates using DD and SLD crossed at seventh or eighth grade.

Finally, 89.6 % of the sample was identified as nondisabled by both the traditional SLD and the 1SD model, and 90.4 % were consistently nondisabled by SLD assessments and the Silberglitt & Gibbons model. Conversely, 1.3 and 1.1 % of the sample was consistently identified as DD and SLD using the above two respective DD approaches.



**Fig. 5.1** Percentage of students identified as dually discrepant and learning disabled (SLD) in each grade level



**Fig. 5.2** Percentage of students identified as dually discrepant and learning disabled (SLD) in each grade level for 1SD and Silbergliitt and Gibbons (2005) models

## Potential Implications

The current chapter examined how well DD criteria differentiated reading skills of children who were below the 25th percentile for reading and who were identified as dually discrepant or non-discrepant. All four approaches significantly differentiated reading skills with mostly large effects. These results were consistent with previous research which also found significant differences in skills between DD and non-DD students (Burns & Senesac, 2005; Fuchs, 2003; McMaster, Fuchs, Fuchs, & Compton 2005; Speece & Case, 2001; Speece, Case, & Molloy 2003). This represents an interesting and consistent finding because the inability to differentiate reading skills of SLD and low-achieving non-SLD students was a common criticism of traditional SLD diagnostic approaches (Aron, 1997). Although the students identified as exhibiting a DD were not necessarily identified as SLD in this example, using DD data for SLD identification could differentiate reading skills better than any previous approach.

The second research question was to describe the prevalence of children identified as needing intensive intervention with DD models. Two of the four DD models had the highest percentage of children identified as needing intense intervention at second grade, but in all models the percentage of children identified as SLD increased in every grade except sixth. Put another way, although second graders were the most likely to need intensive reading interventions, they were less likely to get this help through special education services than their older peers. This could have significant implications for RtI practice because RtI is the use of assessment data to allocate resources most efficiently in order to enhance learning for all students (Burns & VanDerHeyden, 2006) and research has consistently shown that the earlier reading interventions occur, the more successful they are (Clay, 1993; Snow, Burns, & Griffin, 1998). The current data suggest that resources could be allocated to children needing intensive interventions in earlier grades with the DD model than with a traditional SLD wait-to-fail identification model. Moreover, the prevalence rates found here closely approximated the idealized prevalence rate (5 %) of children requiring the most intense intervention in order to implement a sustainable RtI model (Burns & Coolong-Chaffin, 2006; Reschly, 2003).

Using a normative level and criterion-referenced slope (i.e., the Silbergliitt and Gibbons model) has an advantage over other models from a practical standpoint. Problematically, normative slope calculations have typically been conducted in post hoc studies. This is because, if a normative slope is to be calculated for a specific group of children, it must be done after the data are collected. This is not practical from the standpoint of a school wanting to deliver interventions and make eligibility decisions on students throughout the school year. An alternative would be to use a national slope norm, or to use slope norms from previous years for the school or district. But, national slope norms have yet to be established, and using prior years' slope norms may be inappropriate, depending on curriculum and instructional programming changes at that school or district. Using a criterion-referenced slope avoids these problems, as a consistent standard is provided across school years, and

the standard can be known to students and educators on the first day of the school year (Silberglitt & Hintze, 2007). Still, using a normative level allows some control over prevalence rates. When using normative levels high-performing schools can still provide services to the lowest-performing fraction of their student body, as can lower-performing schools, without concern that their performance relative to some benchmark will create unacceptably high or low prevalence rates.

Given that the critical component of RtI is to systematically allocate resources to improve learning for all children (Burns & VanDerHeyden, 2006) and to identify and remediate problems (Christ, Burns, & Ysseldyke, 2005), the goal for RtI implementation is to find solutions rather than failures. Thus, a child who does not benefit from an intervention or instruction should not be thereafter characterized as a “nonresponder,” but the instructional approach should be characterized as insufficient to establish an appropriate and desired response. This is a subtle, but critical distinction. Problems are solved with ecological manipulations and systematic evaluations of those manipulations. The problem is properly described as an interaction between the child’s needs and the services provided. In that sense, labeling the child as a nonresponder does very little to facilitate RtI and has the potential to undermine intervention efforts. Both the right focus (i.e., program effects in terms of student response) and adequate standards for evaluation are necessary for the success of a multi-tiered DD model of RtI.

### ***Future Research***

Although the current data have potential implications for practice, several limitations should be considered. First, these data were from one education district with a relatively homogeneous population. Therefore, replication of this methodology with a more diverse student population would be beneficial. Moreover, this study examined data that were part of a district-wide benchmarking system and were not collected within a fully-implemented RtI model. Therefore, it is unknown how small-group interventions or intensive individual interventions would affect the data. It may also be advantageous to replicate the study with a longitudinal design to examine prevalence of dual discrepancies across grades with a longitudinal design. Finally, slopes of student growth for children receiving intensive interventions are usually collected more frequently than the three benchmarks used here (Reschly, 2003). Future researchers may wish to replicate this study using weekly progress monitoring data.

In addition to replications to address the limitations mentioned above, future researchers may also wish to further examine student and environmental characteristics of children who are dually and nondually discrepant. These data may be more relevant to practice than theory. Therefore, future research in the schools could examine the actual, rather than assumed, implications for resource allocation and SLD identification. Finally, RtI may serve as a defining variable for future DD research. In other words, researchers could implement a sound intervention with children



identified as DD and non-DD to compare growth rates between the two after identification.

## Conclusions

Measuring student competence has long been a frequently researched and debated topic that was intensified with the 2004 RtI provision in federal special education regulations. An operational definition of student response and lack thereof needs to be objective, meaningful, and empirically supported. Traditional approaches to SLD identification were somewhat subjective (Algozzine & Ysseldyke, 1982; 1983; Haight, Patriarcha, & Burns, 2002) and lacked research to support the validity of decisions made with them (Aaron, 1997). Thus, a growing literature already supports the DD approach over traditional SLD identification, but being superior to a highly criticized model is neither convincing nor the goal. Future research and resulting practice should focus on the use of DD, and RtI in general, in the enhancement of student learning and competence with the ultimate yardstick for empirical support being the improvement of outcomes as a result of collecting the data.

Reading is complex and there is certainly more to how well a student reads than how fluently they do so, but ORF is directly linked to comprehension and is an overall indicator of overall reading skills. The data presented here suggest that one of the simplest ways to conceptualize reading assessment could be one of the most useful when identifying students for whom intervention is or is not successful. Additional research is needed, especially given the large number of students who continue to struggle with reading and the promising data presented here.

## References

- Aaron, P. G. (1997). The impending demise of the discrepancy formula. *Review of Educational Research, 67*, 461–502.
- AIMSweb. (2006). *Measures and norms*. Eden Prairie: Edformation.
- Algozzine, B., & Ysseldyke, J. (1982). Classification decisions in learning disabilities. *Educational and Psychological Research, 2*, 117–129.
- Algozzine, B., & Ysseldyke, J. (1983). Learning disabilities as a subset of school failure: The over sophistication of a concept. *Exceptional Children, 50*, 242–246.
- Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research sounding curriculum-based measurement of oral reading fluency decision rules. *Journal of School Psychology, 51*, 1–18.
- Berninger, V. W., Abbott, R. D., Vermeulen, K., & Fulton, C. M. (2006). Paths to reading comprehension in at-risk second grade readers. *Journal of Learning Disabilities, 39*, 334–351.
- Burns, M. K. (2007). Reading at the instructional level with children identified as learning disabled: Potential implications for response-to-intervention. *School Psychology Quarterly, 22*, 297–313.
- Burns, M. K., & Coolong-Chaffin, M. (2006). Response-to-intervention: Role for and effect on school psychology. *School Psychology Forum, 1*(1), 3–15.

- Burns, M. K., Scholin, S. E., Kosciolk, S., & Livingston, J. (2010). Reliability of decision-making frameworks for response to intervention for reading. *Journal of Psychoeducational Assessment*, 28, 102–114.
- Burns, M. K., & Senesac, B. K. (2005). Comparison of dual discrepancy criteria for diagnosis of unresponsiveness to intervention. *Journal of School Psychology*, 43, 393–406.
- Burns, M. K., & VanDerHeyden, A. M. (2006). Using response to intervention to assess learning disabilities: Introduction to the special series. *Assessment for Effective Intervention*, 32, 3–5.
- Burns, M. K., & Ysseldyke, J. E. (2005). Questions about response-to-intervention implementation: Seeking answers from existing models. *The California School Psychologist*, 10, 9–20.
- Burns, M. K., Kwoka, H., Lim, B., Crone, M., Haegele, K., Parker, D. C., Petersen, S., & Scholin, S. E. (2011). Minimum reading fluency necessary for comprehension among second-grade students. *Psychology in the Schools*, 48, 124–132.
- Carnine, D. W., Silbert, J., Kame'enui, E. J., & Tarver, S. G. (2009). *Direct instruction reading* (5th ed.). Upper Saddle River: Merrill Prentice Hall.
- Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review*, 35, 128–133.
- Christ, T. J., Burns, M. K., & Ysseldyke, J. E. (2005). Conceptual confusion within response-to-intervention vernacular: Clarifying meaningful differences. *Communiqué*, 34(3), 1, 6–8.
- Clay, M. (1993). *An observation survey of early literacy achievement*. Portsmouth: Reed.
- Cohen, J. (1989). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum.
- Coles, G. (1998). Reading lessons: The debate over literacy. New York: Hill and Wang.
- Deno, S. L. (1986). Formative evaluation of individual student programs: A new role for school psychologists. *School Psychology Review*, 15, 358–374.
- Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review*, 30, 507–524.
- Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities: Research & Practice*, 18, 172–186.
- Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research and Practice*, 13, 204–219.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal*, 28, 617–641.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5, 239–256.
- Fuchs, L. S., Fuchs, D., & Speece, D. L. (2002). Treatment validity as a unifying construct for identifying learning disabilities. *Learning Disability Quarterly*, 25, 33–46.
- Fuchs, D., Fuchs, L. S., McMaster, K. N., & Al Otaiba, S. (2003). Identifying children at risk for reading failure. Curriculum-based measurement and dual discrepancy approach. In H. L. Swanson & K. R. Harris (Eds.), *Handbook of learning disabilities* (pp. 431–449). New York: Guilford.
- Fuchs, L. S., Fuchs, D., Hintze, J., & Lembke, E. (2006). Progress monitoring in the context of responsiveness-to-intervention. Presentation at the Summer Institute on Student Progress Monitoring, Kansas City, MO.
- Goffreda, C. R., & DiPerna, J. C. (2010). An empirical review of psychometric evidence for the dynamic indicators of basic early literacy skills (DIBELS). *School Psychology Review*, 39, 463–483.
- Good, R. H., Gruba, J., & Kaminski, R. A. (2002). Best practices in using dynamic indicators of basic early literacy skills (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes

- (Ed.), *Best practices in school psychology IV* (pp. 679–700). Washington, DC: National Association of School Psychologists.
- Gresham, F. (2002). Responsiveness to intervention: An alternative approach to the identification of learning disabilities. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 467–519). Mahwah: Lawrence Erlbaum.
- Haight, S. L., Patriarca, L. A., & Burns, M. K. (2002). A statewide analysis of eligibility criteria and procedures for determining learning disabilities. *Learning Disabilities: A Multidisciplinary Journal*, 11(2), 39–46.
- Hasbrouck, J., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for teachers. *Reading Teacher*, 59, 636–644.
- Howe, K. B., & Shinn, M. M. (2002). *Standard reading assessment passages (RAPs) for use in general outcome measurement: A manual describing development and technical features*. Eden Prairie: Edformation.
- Howe, K. B., Scierka, B. J., Gibbons, K. A., & Silbergliitt, B. (2003). A school-wide organization system for raising reading achievement using general outcome measures and evidence-based instruction: One education district's experience. *Assessment for Effective Intervention*, 28(3&4), 59–72.
- Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review*, 37, 374–390.
- McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children*, 71, 445–463.
- Mirkin, P. K., Deno S., Tindal G., & Kuehnle, K. (1982). Frequency of measurement and data utilization strategies as factors in standardized behavioral assessment of academic skill. *Journal of Behavioral Assessment*, 4, 361–370.
- National Reading Panel. (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Reports of the subgroups. Bethesda: National Institute for Literacy.
- Northwest Evaluation Association. (2003). *Technical manual for the nwea measures of academic progress and achievement level tests*. Lake Oswego: Northwest Evaluation Association.
- Northwest Evaluation Association. (2004). *Reliability and validity estimates: NWEA achievement level tests and measures of academic progress*. Lake Oswego: Northwest Evaluation Association.
- Northwest Evaluation Association. (2005). *RIT scale norms for use with achievement level tests and measures of academic progress*. Lake Oswego: Northwest Evaluation Association.
- Reschly, D. J. (2003). *What if LD identification changed to reflect research findings?: Consequences of LD identification changes*. Paper presented at the Responsiveness-to-Intervention Symposium, Kansas City, MO.
- Reschly A, Busch T, Betts J, Deno S, & Long J. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology*, 47, 427–469.
- Riley-Tillman, T. C., & Burns, M. K. (2009). *Single case design for measuring response to educational intervention*. New York: Guilford.
- Roberts, G., Good, R., & Corcoran, S. (2005). Story retell: A fluency-based indicator of reading comprehension. *School Psychology Quarterly*, 20, 304–317.
- Salvia, J., Ysseldyke, J. E., & Bolt, S. (2010). *Assessment* (11th ed.). Boston: Houghton Mifflin.
- SAS (2005). *SAS Version 9.1.3* (software). Cary: SAS Institute, Inc.
- Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment*, 24, 19–35.
- Shinn, M. R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford.

- Silberglitt, B., & Gibbons, K. A. (2005). *Establishing slope targets for use in a response to intervention model (technical manual)*. Rush City: St. Croix River Education District.
- Silberglitt, B., & Hintze, J. M. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23*, 304–325.
- Silberglitt, B., & Hintze, J. M. (2007). How much growth can we expect? A conditional analysis of R-CBM growth rates by level of performance. *Exceptional Children, 74*, 71–84.
- Snow, C. E., Burns, M. S., & Griffin, P. (Eds.) (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academies Press.
- Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology, 93*, 735–749.
- Speece, D. L., Case, L. P., & Molloy, D. E. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research & Practice, 18*, 147–156.
- Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research & Practice, 15*, 128–134.
- Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: review of research. *Psychology in the Schools, 42*, 795–819.
- Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes for two instructional approaches. *Journal of Learning Disabilities, 34*, 33–58.
- VanDerHeyden, A. M., Witt, J. C., & Barnett, D. A. (2005). The emergence and possible futures of response to intervention. *Journal of Psychoeducational Assessment, 23*, 339–361.
- Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly, 45*, 270–291.
- Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S., Chen, R., Pratt, A., & Denkla, M. B. (1996). Cognitive profiles of difficulty-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experimental deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88*, 601–638.
- Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*(2), 85–120.
- Yeo, S. (2009). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31*, 1–12.

## Part II

# Recommendations for Test Developers

An important consideration when using any measure of fluency is the extent to which a set of scores are reliable and valid and lead to appropriate decisions about students. What constitutes a reliable score, however, is dependent on the measurement framework used to estimate reliability coefficients. In Chap. 6 of this section, Christ and colleagues summarize aspects of classical test theory, generalizability theory, and item response theory as mechanisms for establishing the reliability of scores. They discuss Kane's (2013) interpretation of test validity and use argument framework as a relevant model to connect psychometrics with interpretation. Chapter 7 of this book, by Prindle and colleagues, presents the use of item response analysis that uses both speed and accuracy data to estimate a students' score and improve reliability. Comparisons are made to other measurement frameworks to evaluate differential reliability in scores. Chapters 8 and 9 address topics related to further improving the accuracy and validity of scores from fluency assessments. Smolkowski and colleagues provide an introduction to signal detection theory in Chap. 8. The authors delve into the terminology of universal screening with fluency measures and illustrate the important considerations necessary to evaluate and interpret criterion-referenced scores like benchmark levels of student performance. Lastly, Chap. 9 by Santi and colleagues address a critical issue in fluency research: evaluating the equivalence of scores across different test forms of oral reading fluency. The authors present different methods for equating test scores including linear equating, equipercentile equating, and latent variable equating while highlighting trade-offs among the different approaches.

## Chapter 6

# Foundations of Fluency-Based Assessments in Behavioral and Psychometric Paradigms

Theodore J. Christ, Ethan R. Van Norman and Peter M. Nelson

The concept of fluency influenced both theory and practice in education since the 1970s. For the purpose of this chapter, fluency is the accurate performance of behavior per unit of time. As a prominent example of a fluency measure, curriculum-based measurement (CBM) is designed to measure student performance per unit of time. CBM of oral reading (CBM-R) measures the words read correctly per minute (WRCM). CBM and similar fluency measures are useful because they are quick to administer, repeatable, and potentially sensitive to changes in student achievement across time (Deno, 1985, 2003). These characteristics are useful because they facilitate the measurement and evaluation of instructional effects within the frameworks of universal screening and time series idiographic analysis. That is, data are collected repeatedly across time to document student's progress and those data are then evaluated to determine whether an instructional program established sufficient effects. This concept emerged along with CBM as educators grappled with the new requirements of special education law, which required measureable goals as part of an individual education plan (IEP).

Although psychometric theory guides the development and use of most tests, it does not provide easy or obvious solutions to meet the new requirements associated with individual goals and progress monitoring. Psychometric theory places emphasis on the development of measures for theoretical or latent traits. There is substantial emphasis placed on the reliability and stability of measurement outcomes with regard to rank ordering. Historically, there was substantially less emphasis on sensitivity to changes in student performances across time. As a result, many of the tests that emerged from a psychometric paradigm were not sensitive to change.

---

T. J. Christ (✉)  
University of Minnesota, Minneapolis, MN, USA  
e-mail: tchrist@umd.edu

E. R. Van Norman  
e-mail: vann0140@umn.edu

P. M. Nelson  
e-mail: nels6964@umn.edu

That limited their utility to monitor IEP goals. In contrast, assessments that emerged from behavioral theory seemed more applicable. That is, the behavioral assessments were often developed to measure very specific behaviors repeatedly across time. They were most often designed to evaluate the influence of the environment on the individual. Reliability was not a primary issue because there is an expectation that behaviors change as a function of the environment and state of an organism. This contrasts with the psychometric perspective. Behavioral assessments were designed to be sensitive and less stable. As is discussed below, CBM emerged primarily out of the behavioral and idiographic paradigm. Once emerged, behavioral and psychometric paradigms quickly became entangled.

The behavioral and idiographic approach is clearly illustrated in the writing of Stan Deno who led the early conceptualization and development CBM. Prior to his work on CBM, he coauthored *Data-Based Program Modification: A Manual* (Deno & Mirkin, 1977). In that text, he defined the principles which later inspired the development of CBM. These principles include:

1. At the present time we are unable to prescribe specific and effective changes to instruction for individual pupils with certainty. Therefore, changes in instructional programs which are arranged for an individual child can be treated only as hypotheses which must be empirically tested before a decision can be made on whether they are effective for that child (p. 11).
2. Time series research designs are uniquely appropriate for testing instructional reforms (hypotheses) which are intended to improve individual performance (p. 11).
3. Special education is an intervention system, created to produce reforms in the educational programs of selected individuals, which can (and, now, with due process requirements, must) be empirically tested (p. 13).
4. To apply time series designs to (special) educational reforms we need to specify the data representing the “vital signs” of educational development which can be routinely (frequently) obtained in and out of school (p. 14).
5. Testing program modifications (reforms) requires well-trained professionals capable of using time series data analysis to draw valid conclusions about program effects (p. 15).

Deno’s and Mirkin’s (1977) manual proposed a number of measurement systems that did not take hold. It was CBM that later emerged as the measurement system to serve the above principles first proposed in the manual.

In that manual and later writings, Deno illustrated the influence of behavioral assessment and idiographic analysis on the development of CBM (Deno, 1985, 1989, 1990, 2003); however, he also explicitly stated an interest to use psychometric methods to develop and evaluate the qualities of CBM. In his own words, “The measures would have to be [r]eliable and valid if the results of their use were to be accepted as evidence regarding student achievement and the basis for making instructional decisions” (Deno, 1985, p. 221). That was stated as one of the primary values to guide the development of CBM. The use of both the principles of single case time series analysis, along with that of traditional validity and reliability



metrics, prompted the intermingling of behavioral-idiographic and psychometric-nomothetic paradigms. This intermingling has caused confusion over the years, but it also contributed to the uniqueness of CBM as a fluency measure.

The purpose of this chapter is to discuss the previous and future contribution of test theory to the development and evaluation of fluency measures. In our opinion, it is necessary to consider the contexts of both behavioral and psychometric assessment to truly understand fluency-based measures. That context helps establish how classical test theory (CTT) and generalizability theory (GT) remain relevant to the future development of CBM and other fluency measures. The application of these theories will be contrasted with that of behavioral assessment and item response theory (IRT).

## **Behavioral and Psychometric Paradigms**

Assessment is broadly defined as the process of collecting information to make a decision. Two dominant paradigms inform the development and use of assessments in the social sciences. Those are the behavioral paradigm and psychometric paradigm. CBM is a fairly unique example of a standardized assessment procedure that emerged from a behavioral assessment paradigm, yet its researchers and developers quickly adopted a psychometric assessment paradigm (Ardoin, Roof, Klubnick, & Carfolite, 2008; Christ & Hintze, 2007; Deno, 2001). This is likely a source of ongoing confusion because each paradigm relies on distinct assumptions, levels of inference, and sources of evidence. These differences are briefly discussed in the subsections below and include manifest and latent variables, the directness of measurement, differences in response format, repeatability of observations, and the methods by which scores are interpreted. As noted, fluency measures are somewhat unique in that they draw on characteristics of the behavioral and psychometric paradigms.

### ***Manifest and Latent Variables***

One of the fundamental differences between psychometric and behavioral assessment is the attribution and inferences associated with score interpretation. Historically, the psychometric paradigm relied substantially on expectancy or latent trait theory, whereas the behavioral assessment paradigm rejected the need to rely on unobservable theoretical observations or traits. As is discussed below, recent conceptions of validity and test theory may have reduced the necessity to rely on latent traits as part of psychometric test development.

A latent trait is an unobservable characteristic of an individual that explains—or causes—a person's performance. Intelligence and abilities related to reading or mathematics achievement are good examples of latent traits. These traits are often

assumed for purposes of psychometrically based educational assessment. The observed performance on a particular test is inferred to be a manifestation of an underlying trait. As such, observable student performances are described as *manifest*—or observable—and the underlying trait is described as *latent*—or unobservable. The manifest performance is directly observed and the state of the latent trait is inferred. This trait theory is foundational to IRT. In a similar way, expected value theory is the foundation to CTT and GT. Both trait and expected value theories provide that individual observations are mere samples of behaviors that are used to generalize to a larger set of possible observations. Although it is not precisely correct, we use “trait theory” to refer both latent trait and expected value theories.

Despite its integral role in other prominent test theories, trait theory is practically irrelevant to the behavioral paradigm for assessment because behavioral interpretations require low levels of inference. That is, attributions of performance are limited to other observable phenomena rather than a theoretical latent variable. The observable influential variables often include characteristics of the environment, learning history, or observable biology events (e.g., blood pressure, neural activity) of the student. Observable behavior and events are of primary interest in the behavioral paradigm rather than unobservable behavior or theoretical constructs, such as latent traits.

### *Directness of Measurement*

Measurement in the psychometric paradigm is inherently indirect. As described, the purpose of measurement is to estimate the state of a latent trait or average performance. The manifest variables, which are observable behaviors, are indicators of those. These underlying assumptions provide the basis to use multiple-choice items on tests of reading comprehension as indicators of the latent trait (i.e., comprehension) or average performance. The behavior of interest is not circling answers or completing a response sheet, but either could be used as the observed behavior—or manifest variable—that functions as an indicator of the construct of reading comprehension. The same is true for the use of rating-scale responses to gauge other traits or tendencies such as depression. The authenticity of the task is secondary to its perceived value as an indicator of the person’s trait or tendency.

Measurement in the behavioral paradigm is inherently direct. The occurrence and nonoccurrence of the behavior is used as the primary variable of interest. Unobservable variables are of less interest and the consideration of concepts such as reading comprehension or depression requires an operational definition of authentic behaviors. The units of measurement are most often accuracy, frequency, rate, duration, latency, or magnitude of the target behavior. Thus, the issue of directness varies across the two paradigms. Fluency measures are quite common in the behavioral paradigm. They are relatively less common in the psychometric paradigm as they are a direct measure of the number of correct responses by unit of time. Nevertheless, fluency is often considered to be an indirect measure of various academic outcomes. While fluency measures are inherently direct, the manner in which those

measures are applied may be direct or indirect. For example, the number of words read correctly on material drawn from grade-level curriculum can be interpreted as a direct measure of the appropriateness of that curriculum. Alternatively, those same data may be used to make an indirect inference about overall reading competence. The dual purpose of fluency as a direct observation of reading performance and an indicator of reading ability is discussed below.

### ***Power, Speeded, and Fluency Tests***

The two primary modes of assessment within the psychometric paradigm are power and speeded tests. Power and speeded tests are typically constructed in a manner that is “long enough to allow all, or nearly all, examinees to finish” (Crocker & Algina, 1986, p. 145). Even in the case of a speeded test, it is expected that most examinees have sufficient time to respond to all stimuli. The fluency assessments discussed here are timed, continuous performance tasks developed to ensure that all, or nearly all, examinees do not finish and respond to all stimuli. These fluency-type assessments are not always inconsistent with the psychometric paradigm. They are just less typical in the psychometric paradigm.

### ***Repeatability***

Repeatability of observations is of primary importance in the psychometric paradigm. It is assumed that behavior is influenced primarily by a latent trait or tendency, which should be stable or measured reliably. The implication is that repeated measurements should be consistent for the same individual on tasks where performance is influenced by a latent trait or tendency to perform in a similar manner. Inconsistency within an administration occasion or across two measures that are administered in close temporal proximity are both indicators of poor—or less reliable—measurement. In contrast, the reliability of an observation is of secondary importance in the behavioral paradigm. It is assumed that behavior is influenced primarily by proximal and observable events. These events are often related to the environment, but they might also relate to biological events within the individual or learning history. Applied behavior analysis depends on the variation of behavior across observations to glean the functional relation of behavior and other observable events.

Fluency-based assessments are often used in the idiographic tradition of progress monitoring, which is meant to evaluate the specific effect of an intervention on the individual. Ongoing measurement to gauge response to intervention requires assessments to be highly sensitive to relevant changes in student performance over time. Repeated administrations of fluency tests produce estimations of response rate, which are likely to be more variable than the outcomes of power and speeded-tests. This particular feature of fluency-based measures might provide enhanced sensitivity to intervention effects, which is often assumed, but it might also reduce

the psychometric reliability and equivalence of scores collected with alternate measurement forms, occasions, and raters.

### *Idiographic and Nomothetic*

The nomothetic tradition in science aims to describe and explain general principles and laws. In contrast, the idiographic tradition in science aims to describe and explain individual events. As discussed, psychometric assessment tends to converge with the nomothetic tradition, and behavioral assessment tends to converge with the idiographic tradition. Psychometric assessments emerge from sampling theory. Examinees and items are sampled from a larger universe to estimate their characteristics and relations which are tested to evaluate their implications and generalizations. Tests are often developed with samples separate from those persons for whom the tests and test scores are used. The items and behaviors sampled are not necessarily of interest. Instead, they are representative samples of relevant items and behaviors. Tests are then used summatively and the interpretation of the test score is contextualized by the performance of the group, or normative sample. One assessment—rather than repeated assessments—is often appropriate because the nomothetic approach supports the assumption that the best context for interpretation of the stable latent trait or tendency is the group. The status of the test score in the larger universe of other examinees establishes the context for interpretation.

Behavioral assessments emerge from specific circumstances. They are developed as authentic measures of behavior that occur in a specific set of circumstances. They are often developed to address problems or concerns of an individual person, rather than a generic sample of examinees. The individual is then assessed repeatedly within or across conditions. The scores are evaluated for their level, trend, and variability. The stability or variation of behavior for the individual(s) and circumstances are evaluated. Importantly, variation in behavior across observations is often useful within the behavioral and idiographic traditions. That variation in behavior is examined along with covariation of other observable events. This comparison helps researchers and practitioners to test the functional relations between the target behavior and other events. The goal is not to generalize to others as the performance of independent samples or norm groups are rarely the context for interpretation. The individual's performance and, perhaps, that of other individuals in the environment, provides the context for interpretation.

It is useful to contrast these traditions, but they are not mutually exclusive. As described by Kimble (1989), "Every individual is a unique expression of the joint influence of a host of variables. Such uniqueness results from the specific (idiographic) effects on individuals of general (nomothetic) laws" (p. 495). That is, the case of an individual person or circumstance is unique; however, there are likely general principles that apply to an individual as well as to a group. The adoption and use of an idiographic approach does not necessarily discount consideration of general principles. Neither does the adoption and use of nomothetic approach negate the consideration of the uniqueness of specific persons or circumstances.

## Fluency as an Indicator

CBM emerged from a behavioral and idiographic paradigm. The early emphasis was on repeated performance sampling, graphic display of time series data, and rate-based recordings of fluency. At the same time, there was an early emphasis on the psychometric and nomothetic paradigm. This was described by Deno in a very early paper on the development of CBM.

Since the purpose of developing measurement procedures was to place in teachers' hands a simple way to routinely monitor student achievement in the curriculum, a set of design characteristics was specified that guided all research and development activities. The measurements would have to be (1) *Reliable and valid ...* (2) *Simple and efficient ...* (3) *Easily understood ... Inexpensive* since multiple forms were to be required for repeated measurement. (1985, p. 221)

The very first design characteristic elicits the psychometric paradigm. A review of early summaries of psychometric development established that the work was wholly dependent on CTT (e.g., Marston, 1989), which was characteristic of most psychometric development at the time.

In addition to the design principles above, there is also evidence that CBM scores were intended as “indicators” rather than merely using fluency as the behavior of interest. For example, CBM was defined as a general outcome measure such that individual performances were construed as indicators of general academic health and general performance in the annual curriculum (Fuchs & Deno, 1991). The term “indicator” was popular in the early years (Shinn, 1989). Later, CBM was sometimes referred to as a Dynamic Indicator of Basic Skills (DIBS) (Shinn, 1995).

The frequent use of the term “indicator” suggests that the observed behavior was used to infer something more. For example, educators, administrators, and parents infer that the observed oral reading fluency (WRCM) indicates performance in the curriculum and not just their performance on a particular passage at one point in time. Fluency scores are sometimes interpreted as an indicator of generalized reading competence and even reading comprehension (Good, 1998; Shinn, 1992). Those inferences relate to the latent construct of reading achievement and not simply to the observed behavior at a particular point in time. As a result, the validity requires evidence and evaluation of the degree to which scores actually represent the latent trait, tendency to behave, and domain of interest.

To summarize, when individuals collect CBM or similar fluency scores, particularly across time, there is often an implicit assumption that performance is an indicator of a latent trait or broader domain. This inference goes beyond the observed behavior so it should be evaluated. Kane (2013) helps conceptualize the validity of fluency as an indicator of academic well-being. There are at least three assumptions which require evidence and evaluation: (a) the theory is plausible, (b) predictions about observable phenomena are reasonably accurate, and (c) indicators provide appropriate estimates of the construct. These assumptions are discussed in more detail below. In addition, there are other assumptions related to scoring, content, and material development. These issues are also addressed below; however, it is

important to emphasize that CBM materials were initially curriculum-based and curriculum-derived; the stimulus materials were sampled directly from the curriculum. Generally, that is no longer the practice.

There was some research conducted in the 1990s to support the use of curriculum-independent material for evaluating reading fluency (Fuchs & Fuchs, 1992; Fuchs & Deno, 1994). There was also research conducted to examine the variability of student performance across time (Shinn, 1989; Good & Kaminski, 2002). It was observed that curriculum-sampled stimulus materials were often highly variable, which resulted in variable student performance across forms. Those findings spurred the development of standardized instrumentation. Up to that time, only the procedures for administration and scoring were standardized. The actual materials were curriculum based. This illustrates how CBM continued to diverge from the behavioral assessment paradigm, which would be more consistent with curriculum sampling, to a psychometric approach, which would typically require standardized instrumentation.

## **Psychometric Theory and Inference**

Issues of validity are more burdensome and complex for psychometric assessment than behavioral assessment. The burden is greater in the psychometric context because the level of inference is more substantial. The complexity is greater because there are multiple threats to the validity of inferences, which must be identified and evaluated. It requires a number of inferences to move from an oral reading observation to an estimate of reading health and literacy. Test theory is useful to evaluate the veracity of the claim that WRCM functions as an indicator. As will be evident in the subsections below, the consistency of observed performance plays a central role in the psychometric validation of fluency measures; however, the way in which consistency is defined and measured differs across theoretical approaches to validation.

### ***Classical Test Theory (CTT)***

As discussed, during a CBM-R assessment, students read aloud from a grade-level passage. This method—as with any measurement in the social sciences—is always associated with error. The magnitude of error associated with a test score is likely to differ for any number of reasons, not limited to the quality of the measurement tool. Regardless of the magnitude, estimating how much error is associated with scores is requisite for any further score interpretation. CTT is the most long standing framework to evaluate the precision of measurement (Spearman, 1904; Brennan, 2011). The CTT model includes terms for the observed score ( $X$ ), true score ( $T$ ), and a general error term ( $E$ ):

$$X = T + E$$

From the above model, it follows that a student's fluency score ( $X$ ) is equal to the sum of their true score ( $T$ ) and error score ( $E$ ). The true score and error term in the CTT model are unobservable. A student's true score is conceptually equivalent to the average of an infinite number of observed scores. Because this computation cannot actually be completed, researchers rely on some of the assumptions of CTT to estimate reliability. In CTT, it is assumed that the true score and error score are unobservable and uncorrelated. In addition, errors of observations are uncorrelated and test forms (e.g., CBM probes) are assumed to have equal observed score means, variances, and covariances.

In the CTT model, if multiple observations of individuals in a group are consistent across forms, occasions, or administrators, this is taken as evidence that the relation between the observed score and true score is high. If observed scores are inconsistent, it follows that the relation is weaker. Assuming the tendency to behave or respond throughout an assessment remains constant, changes in the observed score ( $X$ ) across conditions are attributed to error ( $E$ ). A strong relation between true and observed scores is evidence for reliability.

Perhaps equally important is the direct connection between reliability and precision—higher levels of consistency allow for more confidence in the observed score as an estimate of the true score. For example, knowing that a student reads 60 WRCM is much more useful if a teacher is confident that the student would perform consistently on other administrations. It would make little sense to interpret the above score if the same student, when given a new reading passage, read 100 WRCM. In this case, the relation between observed score and true score would be much too large for any meaningful interpretation.

In the context of CBM, a student's WRCM may change slightly as a function of different forms, occasions or administrators. The CTT model assumes the construct is relatively stable over any brief period of time so fluctuations in the observed score are attributed to error. For example, the quality of reading passages for CBM is likely to impact—sometimes to a large degree—students' WRCM. Consequently, researchers who develop and disseminate reading passages for screening and progress monitoring must provide information to support score consistency across reading passages, which is indicated by evidence of alternate-form reliability. A similar experimental procedure is applied to students' scores over a short period of time (test–retest reliability). It is important to note that in each case, there is only one error term. Thus, if a researcher designs an experiment to estimate test–retest reliability, he/she attributes all error to the manipulated variable (in this case, time) and assumes that additional error is negligible.

Fluency scores likely reflect a confluence of variables, one of which is the fluency construct, but many of which are not. The potentially large number of factors influencing fluency scores makes the CTT framework less useful for estimating the reliability of CBM. Recall that within CTT, error is assumed to be constant and uncorrelated across measurement occasions. Even if a large degree of control is exercised by researchers, the reported reliability coefficients may not be replicable



in applied settings. In the case of fluency scores, it may be more useful to consider the relative contribution of a variety of error sources that are expected to vary from one measurement occasion to the next. The following section briefly describes GT, which, as an extension of CTT, provides the framework for such considerations.

### ***Generalizability Theory***

Although the CTT framework continues to remain relevant in the measurement literature, the potential advantage of quantifying multiple sources of error in the same model is readily apparent. Like CTT, GT is a statistical approach to estimate the consistency of measurements. Both approaches rely heavily on the notion of domain sampling as both treat items (i.e., passages) as one sample from an infinite domain of such items; however, GT expands on the CTT model by allowing a more particular definition of the measurement process. More specifically, each sample of behavior is considered to be a sample from a *universe for generalization* that consists of all possible observations of the person. Importantly, the researcher or test developer defines the universe for generalization and thus defines which factors are included in the universe and which are not. These measurement factors are referred to as *facets* in GT and together define the measurement conditions that the test developer wishes to generalize across. The identification and measurement of facets is one of the most powerful contributions of GT because it allows the various sources of error that were bound together in CTT to become disentangled and estimated individually. Brennan (2001) illustrates this conceptual difference in the following equation:

$$X = \mu_p + E_1 + E_2 + \dots + E_H$$

Where  $\mu_p$  is the *universe score* and the error terms represent the variance associated with a given set of fixed or random facets of measurement. The universe score is the analogue of the true score in CTT but can be markedly different in meaning depending on the universe of generalization (i.e., number and type of facets) defined by the researcher. From the above equation, it follows that GT is akin to a random effects analysis of variance (ANOVA), with each facet represented by a fixed or random factor.

In addition to the variance components for facets included in the universe of generalization, there are two reliability coefficients of interest in GT, each related to the intended use of the measurement itself. More specifically, GT provides researchers with the means to estimate measurement error for relative decisions and absolute decisions. As a normative metric, relative measurement error provides an estimate of consistency in the rank order of individuals across conditions. This contrasts with absolute measurement error, which estimates the consistency for comparisons against a criterion that is unrelated to the performance of other individuals. In the context of CBM, both absolute (i.e., whether a student is meeting grade level bench-

marks) and relative (i.e., how a student scores relative to his/her peers) decisions are important.

Researchers generally estimate error for relative and absolute decisions using information generated from the random effects ANOVA model mentioned above. During this process, referred to as a “Decision Study” in GT, decision makers may use information about some or all the facets for measurement to estimate reliability. Regardless of the facets included in the universe of generalization, measurement error for relative and absolute decisions is defined differently as different variance components are included in each computation. For relative decisions, only variance components that would be expected to impact students’ relative standing are included for analysis. In the context of CBM research, variance components for items and administrators are omitted from any estimation of relative error if all students respond on each passage and are exposed to each administrator. As a result, the main effect of passage and administrator is constant across students and is not required for estimating the error for relative decisions; however, all two and three way interaction terms are included. The GT coefficient for relative decisions is referred to as the generalizability coefficient and is interpreted as the proportion of observed score variance due to universe score variance.

In contrast to the measurement error associated with relative decisions, error for absolute decisions includes all facets of variance (as defined by the decision maker). The resulting coefficient is referred to as the dependability index and is appropriate for considering an individual’s score relative to a criterion of interest (e.g., grade-level benchmark in CBM) across the conditions included in the universe for generalization.

### *Item Response Theory*

Item response theory (IRT; see Chap. 7 this volume) is the primary alternative to CTT and GT. It is substantially dependent on latent trait assumptions. Whereas the primary measure of interest within CTT and GT is overall test scores, IRT focuses on individual item responses. More specifically, researchers and test developers use IRT to model the probability of a correct response to an individual item through mathematical functions. These functions incorporate characteristics or parameters of an individual item (e.g., discrimination, difficulty, and pseudo-guessing), as well as the estimated level of the latent trait in the individual to predict the likelihood that the examinee will endorse a particular response or answer an item correctly.

Brennan (2011) offers an apt analogy to compare and contrast the main perspectives of measuring performance with CTT, GT, and IRT:

Consider individual items as trees and the universe of items as the forest. If we focus on individuals trees as we do in IRT, then we are easily oblivious to the forest. If we focus on the forest, the trees are indistinguishable. (p. 17–18).

That is, using a CTT or GT framework comes at the expense of losing information at the item level. Conversely, adopting an IRT-based approach—particularly mul-

tiparameter models—restricts traditional analyses for overall test scores. Although CTT and GT remain as relevant and useful psychometric theories, IRT has contributed immensely to the research literature on psychological measurement and has greatly enhanced scale development. One only needs to look at the research on computer adaptive testing for evidence of IRTs contributions (Ware, Bjorner & Kosinski, 2000).

An individual's level of a latent trait or  $\theta$ , is the closest approximation to a test score within IRT; however, the number of items an examinee answers correctly is not necessarily equal to the total score. Instead, responses to individual items are weighted based on their associated parameters. The estimated value of the latent trait in IRT provides the most likely explanation for any given response pattern (hence the name maximum likelihood estimate). Because of the inherent connection to individual items, traditional notions of reliability (i.e., true score variance  $\div$  observed score variance) are not directly applicable to test scores calculated from IRT analysis.

Although there are reliability-like statistics that can be computed from IRT analysis, reliability within an IRT framework is often conceptualized as the magnitude of error variance at a given ability level (Brennan, 2001). This characteristic of IRT analysis overcomes the restriction of equal SEM as well as potential problems with replication of reliability in CTT. The conditional error variance is the inverse of the information function at that ability level. This conceptualization of reliability is not based on semantic interpretations, it is a mathematical derivation (Brennan, 2011). Implicit with the derivation of the standard error of the maximum likelihood estimate of  $\theta$ , there is no consideration of sampling items (Brennan, 2011). If items are not sampled, they are fixed. Because items are fixed, replications within IRT, conceptually, involve the administration of items with identical parameters. That is, conditional standard errors within IRT represent replications over perfectly parallel forms. When researchers limit the role of facets as in IRT, the power of replications becomes limited. Therefore, current applications of IRT do not allow researchers to investigate different sources of measurement error as in GT. The different sources of measurement error, as stated previously, are of high substantive interest. Limiting the role of facets in IRT restricts reliability estimates for scores across different conditions of interest. In essence, the conditional standard error of measurement within IRT is only demonstrative of the reliability of scores across a very narrowly defined set of conditions (i.e., perfectly parallel forms).

Other assumptions associated with IRT pose particular problems for measuring fluency. Researchers have advocated for the application of IRT models to CBM-R passages at a word-by-word level, with words functioning as items and passages functioning as tests (Betts, Pickart & Heistad, 2009). That is, parameters would be derived for each word in the CBM-R passage and the resulting  $\theta$  estimate would be a measure of the student's ability. There are several obstacles associated with the application of IRT to CBM-R. First, most applications of IRT rely on the assumption that the examinee has an unlimited amount of time to complete the measure and each examinee has a chance to respond to every item (i.e., a power test). As stated before,

CBM-R and other fluency-based measures are constructed with the assumption that very few examinees will have the opportunity to respond to every item. Granted, time-based IRT models have been developed (e.g., Wise & DeMars, 2006). Yet it is unclear how such models could be adapted to fluency-based measures.

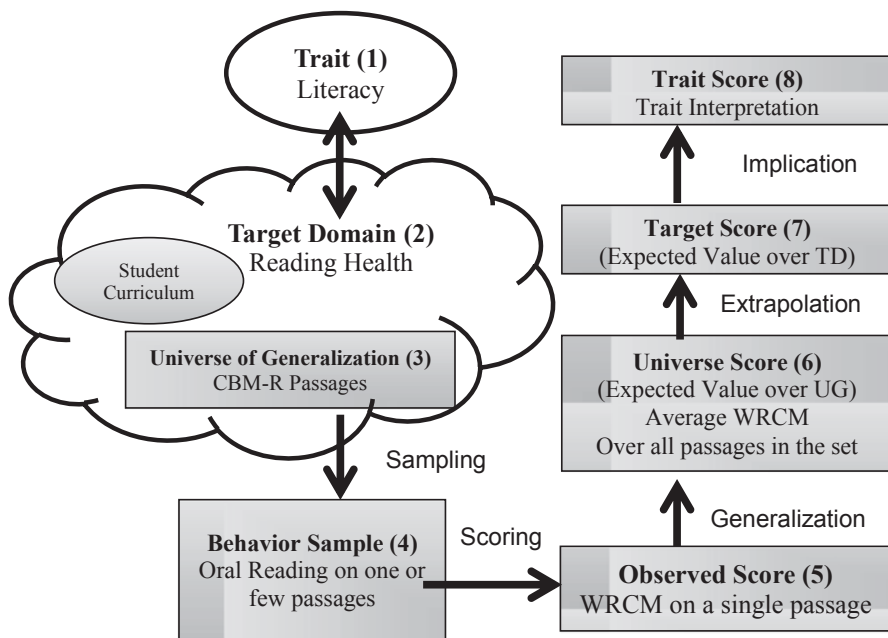
Another issue related to applying an IRT model at a word-by-word level is the assumption of conditional independence. The assumption of conditional independence states that the likelihood of answering one item correctly on a test does not affect the likelihood of answering any other item on the test correctly. The robustness of conditional independence is often debated within IRT (Wainer, Wang, Skorupski, & Bradlow, 2005), but the argument is typically framed as individuals responding to a small cluster of items related to a common passage. The argument is rarely framed as a timed task of reading connected text, often with repeated words.

Finally, it should be noted that the process of deriving item parameters for tests is time and resource intensive. There are no gold standards for sample sizes, but simulation studies suggest that for a 60-item, three-parameter model, over 1000 responses might be needed (Baker, 1992). If one was interested in parameterizing a CBM-R passage for upper elementary students with the same three-parameter model, the recommended sample size grows considerably. Related to the point raised earlier, the issue of parameterizing items when not every examinee is equally likely to respond to every item adds another level of complexity. We do not raise these concerns to be dismissive of the application of IRT to fluency-based measures. Rather, we are intrigued and excited about the application of new scaling techniques to measures. We raise these issues as a caution for viewing IRT as an all-encompassing replacement of CTT and GT.

### ***An Argument-Based Framework***

Before placing too much stock in the meaning of reliability—whether derived from CTT, GT, or IRT—it is helpful to carefully consider the broader context and purpose for reliability research. In the case of CBM-R, teachers are tasked with making meaningful decisions about students' literacy skills using a timed measure of WRCM. In a broad sense, those who create an assessment must compose an argument for its use in applied settings. Some of this argument is grounded in traditional notions of reliability and other parts related to validity, but the argument should be a coherent and defensible narrative. Kane's (2006, 2013) perspective on validation incorporates aspects of both reliability and validity, creating a framework to guide test development and evaluation. Kane's interpretation and use argument (IUA) framework is particularly relevant here because it illustrates the connection between concepts that are sometimes considered distinct and draws attention to the subjective nature of the psychometric argument itself. Kane describes an interpretation and use approach to validation. This contemporary approach to validity unifies the various aspects of test development and test score evaluation that are discussed previously in this chapter.

## Hypothesized Empirical Relations



**Fig. 6.1** Depiction of hypothesized relations between the latent trait of literacy and the target domain of generalized reading health

There are at least three assumptions in Kane’s IUA that apply to the use of fluency measures as indicators (Rodriguez, personal communication, 2013). These include: (a) the theory is plausible, (b) predictions about observable phenomena are reasonably accurate, and (c) indicators provide appropriate estimates of the construct. The IUA includes at least three inferences: (a) scoring (the scoring rule provides clear, consistent scores), (b) generalization to the universe of generalization (passage sampling, behavior sampling, rater sampling), and (c) interpretation from the value of the indicator to the value of the construct (see Chap. 13, this volume). We use Kane’s conception of IAU to provide a context and perspective on test development. In the case of fluency measures, it is very helpful to apply CTT and GT to spur development, evaluate the assumptions, and test the inferences. A generic depiction of the inferences and components associated with interpretation are presented in Fig. 6.1.

### Trait or Tendency (1)

As discussed, the latent trait (IRT) or behavioral tendency (CTT, GT) influences or causes performance in a particular domain. In our example, the trait is literacy and

the target domain is generalized reading health. Figure 6.1 illustrates that performance in the target domain is influenced by the latent trait.

### **Target Domain (2)**

As illustrated by the cloud-shape in Fig. 6.1, target domains are often difficult to define and even more difficult to observe because they are diffuse and often substantial in size. An exhaustive set of observations, in all possible contexts, is rarely possible. For example, it would be very difficult to observe a student's performance on all reading activities encompassing all reading attributes and behaviors. There are many aspects of reading to consider, which might include self-regulation along with phonological, orthographic, and semantic skills. At the same time, there are many forms of reading stimuli to consider, which include websites, books, newspapers, magazines, menus, and instructional materials. Each type of stimuli might vary in length and content. They also might vary in syntactic, semantic and orthographic complexity. In sum, the target domain is large. It requires sampling rather than comprehensive observation of the whole domain. It is necessary to define a subset of activities and materials that are intended to represent the target domain for assessment purposes. In the language of GT, that is the universe of generalization.

### **Universe of Generalization (3)**

As discussed in the subsection on GT, The universe of generalization defines the characteristics of all possible behavior samples and stimulus materials that might be used for assessment. Note that the universe of generalization is entirely defined by researchers and test developers. For example, the universe of generalization for CBM-R might be restricted to oral reading of narrative-type grade-level passages. This definition substantially restricts the variety of possible behaviors and observations associated with the much larger target domain. It excludes alternative behaviors, which might include responses to multiple-choice comprehension items. It excludes passages that are out of grade level. It also excludes informational or poetry passage types. In this example, only a small portion of the target domain is included in the universe of generalization. This restricts characteristics of assessment and qualities of observation. A restricted universe of generalization increases consistency of measurement, but it might also limit the extrapolation of performance on the assessment to performance in the larger domain. At this point, we tend not to think of this restriction as sampling from the target domain—we tend to think of it as restricting the target domain to a testable domain (universe of generalization) only including those tasks and contexts which are assessable. From the universe of generalization, we hope to sample tasks and contexts for any given assessment. Here we run the risk of construct underrepresentation, but realize that practical con-

straints prevent us from testing some aspects of the target domain (so it is less about sampling and more about purposive reduction). This is discussed below.

#### **Behavior Sample (4): Sampling**

The behavior sample is 1 min of oral reading on one of many grade-level narrative texts that are available from the universe of generalization. The data that are provided to the examiner represent oral reading behavior at a particular time and setting and with a particular passage. Only a small portion of the universe of generalization is observed. It is necessary to validate the inference that the limited sample is a quality indicator of the larger universe. In the context of GT, the assumption here is that the tasks are sampled from the universe of generalization, such that they are exchangeable. GT uses the ANOVA framework, which assumes random sampling. We cannot, of course, randomly sample from the universe because it is difficult to define completely and often impossible to observe to a large extent.

#### **Observed Score (5)**

The observed score is the numeric value that characterizes the behavior sample, which requires the development and application of score rules. The examiner must apply scoring rules to derive the unit of measurement that is WRCM for CBM-R. The quality of the scoring rules influences the quality of inferences that follow from assessment outcomes. Interpretation and use of test scores requires an inference that the scoring is appropriate, standardized, and free from bias. Many of the fluency-based score rules are fairly objective and easy to implement, which makes the inference easier to evaluate. Nevertheless, GT provides useful conceptualization and procedures to evaluate the scoring inference, which often includes estimation of variance components associated with the universe of raters, conditions, and procedures.

After applying administration and scoring rules, the observed score is available for the sample observation. Interpretation is rarely limited to the observed score. That is, interpretation is rarely limited to student performance for one setting, passage, administrator, time of day, and day of week. Instead, there is an inference of generalization to the universe score.

#### **Universe Score (6): Generalization**

The universe score is the expected level of performance across all observations in the universe of generalization. This is often operationalized as the average of all possible observations. Figure 6.1 illustrates that the observed score—WRCM—is generalized to represent the expected WRCM in the entire universe of generaliza-



tion, which again, is defined by those responsible for validating the assessment. Both CTT and GT provide methods to examine the generalization inference. Yet, that is still not the inferential conclusion. Up to this point, the inferences apply both to observable attributes and indicators (as to this point, indicators are operationalized functionally and serve the role of observable attributes). Now, we hope to use them as indicators of a construct and thus require deeper inferences and evidence. There are two more likely inferences.

### **Target Score (7): Extrapolation**

The target score is the expected level of performance across all observations in the target domain. In the CBM-R example, the target score depicts generalized reading health, which is a theoretical value. As previously discussed, the target domain encompasses all of the various forms of reading stimuli that vary in length and complexity—most of which were not observed. In this example, WRCM on a CBM-R passage from the universe of generalization is extrapolated to infer the state of generalized reading health.

### **Trait Score (8): Implication**

Finally, the state of the trait is inferred through implication. That is, an estimated level of performance in the target domain is used to implicate the state of the trait, which is literacy in the CBM-R example. This is the ultimate inference, from the value of the indicator to the value of the construct. The assumptions required to support this inference are that (a) the theory is plausible, (b) predictions about observable phenomena are accurate, and (c) CBM as an indicator is an appropriate estimate of the construct. These might lead to the strongest challenges for the use of CBM as an indicator. This is because we must address the challenges of construct underrepresentation and construct-irrelevant variance (Rodriguez, personal communication, 2013).

While the early components of Kane's framework address issues of replication (i.e., reliability), the meaningfulness of extrapolations to the domain and trait is closely aligned with traditional conceptualizations of validity (e.g., content and criterion-related validity). From Fig. 6.1, it follows that Kane's account of the validity argument bridges the interpretive gap between issues of reliability and validity that are sometimes divided from one another; however, more relevant to the current chapter, it also highlights the direct connection between the universe of generalization and the universe score.

Importantly, the universe score only offers reliability evidence for replications conducted under conditions similar to those used to generate reliability evidence. For example, if researchers use GT to account for the effect of various administrators, forms, and occasions, the reliability estimate is only adjusted for those error sources.

This adjustment is further dependent on whether these sources are treated as fixed or random. In this way, researchers exercise direct control over reliability—it is not the measure itself that is reliable, but the scores under a set of prespecified conditions. It follows that if the universe of generalization is small, score reliability will be higher (and vice versa).

## **Aligning Psychometric Theory with Applied Practice**

The manner in which the universe of generalization is defined is perhaps the most important consideration for those who continue to conduct or consume research on fluency measures. Given the instrumental and contextual fluctuations that occur when measuring fluency, it is meaningful to consider the scope of reliability estimates. Assuming that an applied measurement condition (e.g., a student reading aloud from a grade-level passage in the hallway of a school) differs from the conditions used in reliability research, the reliability of the observed score is less predictable.

Previous CBM work using GT addresses some of the most typical reliability concerns—namely differences across forms and raters—but it may be useful to expand the universe of generalization to include variations common in school settings. For example, student characteristics, such as motivation and self-efficacy, may interact with CBM materials and procedures enough to produce unexpected differences in performance. Such fluctuations may be explicitly modeled if included in the universe of generalization. Likewise, it may be useful to expand the universe of generalization to include variations in the life experience (e.g., culture) of students as these factors may be predicted to influence the reliability of CBM-R scores. Given the increasingly diverse population of students who interact with CBM-R materials, novel extensions to the universe of generalization may more accurately reflect the nature of CBM-R practices in schools. Finally, in a recent review of CBM research, Tindal (2013) calls for reliability studies to include aspects of instruction and intervention. That is, the type and intensity of intervention in particular may also influence the reliability of progress monitoring data.

Reliability is largely under the control of those who choose the methods and procedures for its estimation. While it may not be feasible for researchers to capture all possible sources of error in the context of CBM-R, it is worthwhile to closely consider the manner in which the CBM-R is used in schools and think critically about which factors should be subject to empirical review. Recent work examining the reliability of CBM-R demonstrates the potential benefit of such review; however, it may be beneficial to expand the universe of generalization even further to better reflect applied practice. Such an expansion may increase the relevance and utility of CBM-R in school settings and strengthen the validation argument for CBM-R as a tool for monitoring student response to intervention.

## Summary and Conclusion

CTT and GT remain relevant to the ongoing development and evaluation of fluency assessments. That is, CTT and GT rely on expected value theory, which is the basis of many statistical applications (e.g., ANOVA). In contrast, IRT is substantially dependent on latent trait theory. We treated these theories as similar in this chapter, but the former assumes a probabilistic relation between one observation and other possible observations. IRT depends on a latent trait, which is unobservable. Second, fluency scores are often reported on their natural scale, which is generally consistent with behavioral assessments. This works well with CTT and GT, but IRT sacrifices the natural scale and requires rescaling. Some have criticized that IRT is really more of a scaling technique and less of a psychometric theory (Brennan, 2001).

Although CTT and GT are sample dependent, development and evaluation requires relatively small sample sizes. This places their application within the hands of many researchers and practitioners in education. IRT requires large sample sizes, more sophisticated analysis, and often times requires more expensive and difficult to use software. IRT should be evaluated for application with fluency assessments, but short-term progress and wide-scale development by non-psychometric researchers and practitioners is likely to begin with CTT and GT. As we emphasized throughout the chapter, these remain relevant test theories notwithstanding the emergence and recent popularity of IRT in many large-scale testing programs. Many of those IRT applications are more focused on estimating stable traits and tendencies. In contrast, fluency assessments are often used to estimate the level, trend, and variability of student performance within and across time. This generalization requires a highly sensitive scale for assessment and it often also depends on behavioral and idiographic paradigms, which are most consistent with GT and, perhaps, CTT.

Finally, Kane (2006, 2013) provides a useful framework for test development and validity, which is presented in IUA. Kane's framework is useful to help us identify the inferences and assumptions that are inherent to our interpretation and use of assessments. We must evaluate and devise evidence to support or refute those interpretations and uses. CTT and GT provide a very useful set of tools for those purposes.

1. *Trait*: person characteristic that influences or causes performance in the domain
2. *Target domain*: the possible behaviors, conditions, and observations influenced by the trait
3. *Universe of generalization*: subset of possible behaviors, conditions, and observations that are intended to represent the target domain for purposes of assessment
4. *Behavior sample*: a more narrow subset of actual behaviors, conditions, and observations that are assessed

5. *Observed score*: numeric value that characterizes the behavior sample or test performance
6. *Universe score*: expected level of performance across all observations in the universe of generalization (3; e.g., average of all possible observations), which is inferred and not directly observed
7. *Target score*: expected level of performance in the target domain (2), which is inferred and not directly observed
8. *Trait score*: expected level of the target trait (1), which is inferred and not directly observed

## References

- Ardoin, S. P., Roof, C. M., Klubnick, C., & Carfolite, J. (2008). Evaluating curriculum-based measurement from a behavioral assessment perspective. *Behavior Analyst Today, 9*, 36–49.
- Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology, 47*, 1–17.
- Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*(4), 295–317.
- Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*, 1–21.
- Christ, T. J., & Hintze, J. M. (2007). Psychometric considerations of reliability when evaluating response to intervention. In S. R. Jimmerson, A. M. Vanderheyden, & M. K. Burns (Eds.), *Handbook of response to intervention* (pp. 93–105). New York: Springer.
- Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt Rinehart and Winston.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.
- Deno, S. L. (1989). Curriculum-based measurement and alternative special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 1–17). New York: Guilford Press.
- Deno, S. L. (1990). Individual differences and individual difference: The essential difference of special education. *The Journal of Special Education, 24*, 160–173.
- Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*, 184–192.
- Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Reston: Council for Exceptional Children.
- Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional children, 57*, 488–500.
- Fuchs, L. S., & Deno, S. L. (1994). Must instructionally useful performance assessment be based in the curriculum? *Exceptional Children, 61*, 15–24.
- Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*, 45–58.
- Good, R. H., & Kaminski, R. (2002). *Dynamic Indicators of Basic Early Literacy Skills 6th Edition (DIBELS)*. Eugene, OR: Institute for the Development of Educational Achievement. <https://dibels.uoregon.edu>.
- Kane M. T., (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.), Westport: American Council on Education/Praeger.

- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.
- Kimble, G. A. (1989). Psychology from the standpoint of a generalist. *American Psychologist, 44*, 491–499.
- Marston, D. B. (1989). Curriculum-based measurement: What it is and why we do it. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York: Guilford Press.
- Shinn, M. R. (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford Press.
- Shinn, M. R. (1995). Best practices in curriculum-based measurement and its use in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology-III* (pp. 547–567). Washington, DC: National Association of School Psychologists.
- Spearman, C. (1904). The proof of measurement of association between two things. *American Journal of Psychology, 15*, 72–101.
- Wainer, H., Wang, X. A., Skorupski, W. P., & Bradlow, E. T. (2005). A Bayesian method for evaluating passing scores: The PPoP curve. *Journal of Educational Measurement, 42*, 271–281.
- Ware, J. E., Bjorner, J. B., Kosinski, M. (2000). Practical implications of item response theory and computer adaptive testing. *Medical Care, 38*, 73–82.
- Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*, 19–38.

## Chapter 7

# Using Response Time and Accuracy Data to Inform the Measurement of Fluency

John J. Prindle, Alison M. Mitchell and Yaacov Petscher

The purpose of assessment is to collect information to estimate a person's true ability in a performance domain. One might have a student read for 60 s with the aim of using that 1-min sample to approximate that student's likelihood of being successful at reading and understanding grade-level text. Assessments must have high levels of validity and reliability to target the appropriate performance skills efficiently. Although most assessments focus on accuracy, considering a student's pattern of correct versus incorrect responding, not all capture speed of response as a performance factor. Leveraging all extant data to increase reliability in estimation of scores is a potentially beneficial effort for those in both practice and research settings.

Due to the presence of technology in educational settings, computerized assessments have become increasingly more common (Gray, Thomas, & Lewis, 2010; Miranda & Russell, 2011; Pressey, 2013). One benefit of computerized testing is the built-in capacity to capture speed of performance in addition to accuracy. Computers allow for the precise logging of item response time without being contingent on time as a limiting factor, as an alternative to pencil–paper fluency tasks. Specifically, response speed on an item can be captured even when students' are not limited to a certain timeframe within which to respond, presenting a broader context for use of this information (Petscher, Mitchell, & Foorman, 2015). Although one might posit that this additional information would be beneficial, this question

---

J. J. Prindle (✉)  
Max Planck Institute, Berlin, Germany  
e-mail: prindle@mpib-berlin.mpg.de

A. M. Mitchell  
Lexia Learning, Concord, MA, USA  
e-mail: amitchell@lexialearning.com

Y. Petscher  
Florida Center for Reading Research, Florida State University,  
Tallahassee, FL, USA  
e-mail: ypetscher@fcrr.org

requires empirical validation. Because time and precision are both premiums in the world of assessments, our goal in the chapter is to highlight how classical test theory (CTT) and item response theory (IRT) compare to each other in terms of accuracy data and then how each of these approaches differ when response time is used. Our chapter begins with a discussion of fluency and how it is typically measured, followed by a quick treatment of the importance of response latency in the literature, and how vocabulary knowledge could be used as an outcome with attributes of speed and accuracy. Next, we focus our attention on measurement models with details on distinct aspects of classical test and item response theories and introduce the conditional item response theory model (CIRT) for speed and accuracy. The chapter concludes with an empirical example and considerations for future directions. Note that this chapter serves as an extension and replication of a similar work found in Petscher, Mitchell, and Foorman (2015).

## Measures of Fluency

Fluency is frequently defined as a composite of speed and accuracy in performance (Chard, Vaughn, & Tyler, 2002). Fluency measures are used in multiple academic domains, including reading, math, and writing. In the domain of reading, fluency tasks have been used to measure knowledge of letter names, letter sounds, single words, connected text, or gist-level comprehension (e.g., maze tasks). Oral reading fluency, which measures a student's rate of reading of connected text, demonstrates a high correlation with measures of reading comprehension, making it a common measure used for screening students for risk of reading difficulty (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Assessments of reading fluency are generally found in two forms: those that assess student skills using word or object lists (e.g., letter sound fluency) and those that measure reading using connected text (e.g., oral reading fluency).

Curriculum-based measurement (CBM) is a form of norm-referenced fluency assessment that can be used multiple times a year to measure growth in performance in academic skills (Deno, 2003). Relative to many other types of standardized assessments, this method is quick to administer and cost-effective while still supplying reliable data that reference individual student performance compared with a normative sample. These tasks often also demonstrate greater sensitivity to change in specific skills than global standardized achievement tests, making them more appropriate to administer in frequent intervals. CBMs are frequently used to screen students for risk of failure in an academic domain (Cummings, Atkins, Allison, & Cole, 2008) and to monitor student's growth in a skill over the course of the year (Fuchs, Deno, & Mirkin, 1984). Further, an integral aspect of fluency scores relates to measuring the speed of performance. In these CBM tasks, a student may be given a worksheet of subtraction questions and asked to answer as many questions as possible in 3 min or presented a narrative passage and instructed to read for 1 min. The student's rate of performance is determined by looking at their number of correct responses or words read accurately within the given time limit (Deno, 2003).



It is well known that most fluency assessments are based in CTT (see Christ, Van Norman, & Nelson, Chap. 6, this volume). A fundamental assumption of CTT, as it pertains to reliability, is that the standard error of measurement is assumed to be constant across the range of scores in a population. Despite this assumption, Poncy, Skinner, and Axtell (2005) applied generalizability theory to CBM probes and found that reliability ranged from .81 to .99. These results indicate that, for some individuals, fluency scores were highly accurate and reliable; for others, fluency scores were observed to have relatively lower reliability. This result has been replicated in other studies pertaining to fluency, suggesting that standard error is not equal across a population (Christ & Silbergitt, 2007; Mercer et al., 2012). While a score for one student may be highly reliable, another student's score may not have sufficient reliability. Such variance in reliability may have meaningful implications for the validity of individual student's scores, even if the population reliability is deemed to be sufficient. One core feature of CBMs is the presence of multiple equitable assessment forms, such that progress can be monitored across the year using the same baseline. Although CBMs have generally acceptable reliability across forms, concerns have still been documented related to the presence of form effects (e.g., Cummings, Park, & Schaper, 2012; Francis et al., 2008; Petscher & Kim, 2011). If two passages are not adequately equated in terms of difficulty, misinterpretations may be made about a student's growth in a particular skill or their response to specific interventions that they receive.

Overcoming the reliability concerns resulting from static standard error levels or form effects, while still allowing consideration of speed as a factor, may require an alternative framework to CTT. CBMs allow for one way to capture both accuracy and speed factors at a global level, as a student's accuracy level on a number of items is determined within a specified timeframe. Computerized assessment systems provide the opportunity to capture item response time information without reliance on time contingencies. In this testing paradigm, students are not limited to a certain timeframe within which to respond. It is important to acknowledge that utilization of computerized testing would not replace the practical benefits of time-limited CBMs, such as speed, convenience, and relatively intuitive administration and interpretation. But by utilizing the response information gathered via computerized assessment we may be able to reconsider what constitutes "fluent performance" and overcome the dichotomous choice identified by Cattell (1948) between tests that serve to measure either power (i.e., accuracy given unlimited time) or performance (i.e., ability given limited time).

## Capturing Response Latency

Speed of response, also termed response latency, is considered an impactful factor in performance across a number of different domains, including cognitive science (Sternberg, 1969) and psycholinguistics (Goodglass, Theurkauf, & Wingfield, 1984). The significance of response latency was established in one study where

students were designated as skilled or less-skilled comprehenders (Perfetti & Hogaboam, 1975). Low comprehenders had longer vocalization latencies for words they were shown and asked to pronounce, especially for low-frequency words or pseudowords. In fact, the skilled and nonskilled readers demonstrated similar response times when presented with high-frequency words; however, skilled students were faster in responding to low-frequency words and pseudowords. These results provide a seminal demonstration of the significance of response time, beyond accuracy alone, in providing a comprehensive measure of performance.

One factor to consider when recording response time is the nature of the testing format. Measures structured in lists or maze tasks that are limited to a few sentences are likely suited to item response accuracy modeling, because the individual items can be programmed in a computer environment where response accuracy and response time could be captured accurately. On the other hand, connected text measures may not be as easily incorporated into current item response models because the data would violate the assumption of local item independence due to the repetition of many high-frequency words in a given text. Further, accurately obtaining response time per word would likely be challenging given that you would need a way to measure time spent on each word in a sequence, which would likely introduce threats to reliability. Thus, in order to incorporate response time into a measurement model, attention must be directed to the manner in which questions are delivered.

## Measurement Model Considerations

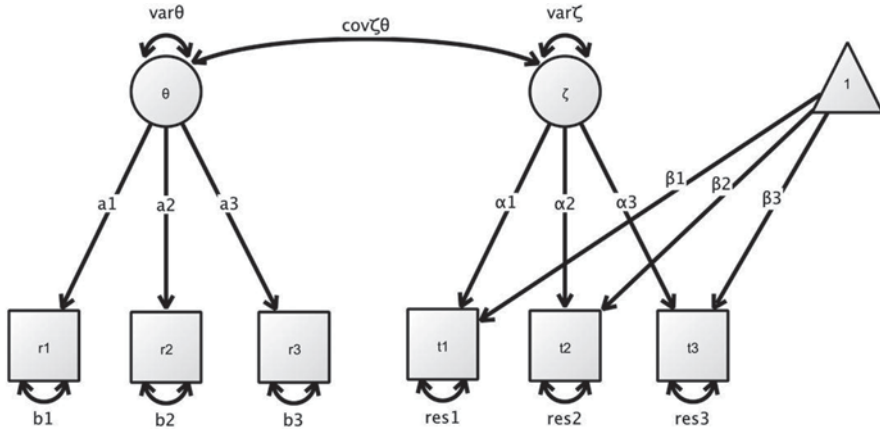
As noted previously, the psychometrics of scores from CBM assessments are often rooted in CTT. Common CTT terms for reliability include internal consistency, test-retest, parallel-form, and split-half reliability. At its core, CTT relates an observed score ( $X$ ) to a true score ( $T$ ) and random error ( $e$ ). The ratio of the true-score variance to observed-score variance is reliability, which separates out the random-error variance from the variance in scores attributable to the ability of the individuals. In addition, there is an inherent assumption of CTT that the standard error of measurement for the test does not vary across a population. Also a limitation, the total test score cannot be directly compared to the difficulty (i.e.,  $p$ -value) of the items in the assessment. A  $p$ -value ranges from 0 to 100%; where low values reflect difficult items and high values indicate easy items. The  $p$ -value of an individual item and the total test scores are on different metrics, thus, it is difficult to make an explicit link between the two. For example, items may range in difficulty from 0 to 100%, while the individual total scores may appear in a raw metric (e.g., 0–26 for a letter naming task), an age-standardized metric (e.g., mean of 100 and standard deviation of 15), or a developmental-standardized metric to capture growth over time (e.g., mean of 500, standard deviation of 100). The relation between an individual's standard score of 100 on a letter naming task and a particular letter's difficulty (e.g., Z;  $p$ -value for the sample=40%) can be quite challenging to associate given this difference in item and person metrics.

An alternative measurement framework, IRT overcomes the limitations of CTT. Its estimation framework is such that it places the item difficulty and the person's ability on the same scale (i.e., a  $z$ -score) via a function known as the item characteristic curve (ICC). The assumption of equal measurement error for all individuals in CTT does not exist in IRT, thus, the reliability of scores is allowed to vary across students (Embretson & Reise, 2000). Two key parameters are embedded within the ICC. First, the item difficulty represents the point on the curve where the probability of correctly answering the question is .50, and is known as the  $b$  parameter. Second, the discrimination represents the steepness of the slope of the ICC, and is known as the  $a$  parameter. Item difficulties typically range from approximately  $-3$  to  $3$  and can be interpreted similarly to  $z$ -scores. Negative values indicate that items are easier while positive values denote harder items. The metric of item discriminations in IRT is  $-\infty - +\infty$ , with optimal values ranging from .8 to 2.5 (de Ayala, 2009); this parameter is related to the notion of the item-to-total correlation whereby large values for item discrimination or the item-to-total correlation suggest a strong relation between the item and the measured construct.

Like the  $b$  value, the ability of the individual, known as a  $\theta$  score, ranges from approximately  $-3$  to  $3$ . With the item difficulties and individual abilities on the same scale, a link can be made between the two. An examinee with an average ability score (i.e.,  $\theta=0$ ) who is presented with an item of average difficulty (i.e.,  $b=0$ ) has a 50% chance of correctly answering that item. When controlling for the difficulty of the item at  $b=0$ , individuals with an ability greater than 0 have greater than a 50% chance of answering the question correctly because their ability exceeds the difficulty of the item. Similarly, individuals with an ability less than zero have less than a 50% chance of answering the item correctly because their ability is lower than the difficulty of the item.

Although item statistics in IRT framework are typically comprised of the item difficulty, discrimination, and a pseudo-guessing parameter, other models have been developed that capture additional sources of variance that may influence an individual's likelihood of correctly answering an item. Testlet effects (Wainer, Bradlow, & Wang, 2007), which describe sets of items that are administered as a bundle (e.g., such as in reading comprehension), are known to influence the probability of a correct response. Item response models have been developed that account for testlet effects; and, as it pertains to the present chapter, it is possible to view item response latency, or the amount of time it takes a student to respond to an item, as another parameter that could influence the probability of a correct response.

With the increasing access and utilization of technology in the classrooms and schools (Blackwell, Lauricella, Wartella, Robb, & Schomburg, 2013; Miranda & Russell, 2011; Pressey, 2013), computerized testing has become a common way to administer academic achievement assessments. Along with recording item-level accuracy data, computers possess the valuable capability to accurately record item-level response times. By recording both time and accuracy, one is able to overcome the dichotomous distinction made by Cattell (1948) who noted that tests function to measure either power (i.e., accuracy given unlimited time) or performance (i.e., ability given limited time).



**Fig. 7.1** A structural equation model diagram for joint response and response time modeling. Ability ( $\theta$ ) and speed ( $\zeta$ ) are indicated by responses and response times, respectively. The covariance between ability and speed ( $\sigma_{\theta\zeta}$ ) is included to indicate that speed may relate to ability scores

### Response Time Models

Though several different theoretical models have been reported in the literature for modeling response time (Schnipke & Scams, 2002), we focus on three that are relevant to the current work. van der Linden and Krimpen-Stoop (2003) proposed that response time is modeled directly into a traditional IRT model, as an interaction between the parameters of response time and accuracy. Specifically, the authors posited that more difficult items are related to longer response times. Prior to the work by van der Linden and Krimpen-Stoop (2003), Scheiblechner (1985) proposed a different method, which stated that the distribution of response time is independent of the item accuracy. This theory is limited because it ignores the ability of the individual. Subsequently, the joint relation between speed and accuracy is unknown. A third method, proposed by van der Linden (2007) is called the CIRT model. This model postulates that item responses vary due to two hierarchical levels, a person/item level and a population/domain level (Fig. 7.1).

At level 1 (the individual level) there are two estimated vectors, one for the individual's item responses (i.e.,  $U_{ij}$ ) and one for the individual's response time (i.e.,  $T_{ij}$ ). The item response vector is defined as:

$$U_{ij} \sim f(u_{ij}; \theta_j, a_i, b_i),$$

where  $u_{ij}$  is the item response on item  $i$  for person  $j$ ,  $\theta_j$  is the latent ability of person,  $a_i$  is the item discrimination, and  $b_i$  is the item difficulty. This function is solved by a traditional two-parameter probability function. The response time vector includes new parameters into the item response theory framework with:

$$T_{ij} \sim f(t_{ij}; \tau_j, \alpha_i, \beta_i).$$

where  $t_{ij}$  is the response time on item  $i$  for person  $j$ ,  $\tau_j$  is the average speed of the individual,  $\alpha_i$  is the discrimination parameter (i.e., variation in response times across items), and  $\beta_i$  is time intensity for the item. Among the response time, speed, and time intensity parameters, only response time is a measured variable. Time intensity is the amount of labor required by the item, or as the effect of an item on the mean log time. Time intensity and the average speed of the individual are considered to be latent constructs. The relation among response time, speed, and time intensity can be expressed as a ratio, where response time is the ratio of time intensity and average speed (van der Linden, 2011). Further, the association between time intensity and average speed is analogous to the interpretations previously described concerning the relation between item difficulty and individual ability. Just as an individual has a higher probability of a correct item response when their ability exceeds the difficulty of the item, so it is also the case that it is more beneficial for the individual's speed to exceed the intensity of the item (i.e.,  $\tau_j > \beta_i$ ) than the reverse ( $\tau_j < \beta_i$ ).

The population portion of Fig. 7.1 (level 2) represents the estimation of the person and item components, as well as the covariances among them, as a function of the level-1 components. The CIRT model includes elements of the theories posited by both van der Linden and van Krimpen-Stoop (2003) and Scheiblechner (1985), whereby, the CIRT approach incorporates the theoretical notion of independence at the individual level of estimation but then includes the joint relation between speed and accuracy at the population level.

Because CIRT is based in IRT, CIRT difficulty and discrimination parameters in the accuracy portion of the model can be compared to those estimated in traditional IRT models. Recent research has suggested that accounting for response time yields more precise estimates of ability due to the joint estimation of accuracy and response latency at level 2 of the CIRT model (Ferrando & Lorenzo-Seva, 2007; van der Linden, 2007). Using the CIRT model has the potential for greater understanding of cognitive processes relative to the task. For example, it may be possible to evaluate the extent to which ability is related to speed, whether difficult items are the most time intensive, and whether other aspects of the test might moderate the relation between an individual's ability and their speed.

## Why CIRT?

A chief concern in educational assessment is maximizing information gained while minimizing time spent in the process of testing. Assessment efficiency is aided when all information gathered can be utilized to inform predictive estimations. Inclusion of speed as a variable beyond accuracy is one way to leverage extant performance information. Although traditional fluency assessments, such as CBMs, capture speed by limiting the response time window, their reliance on multiple forms and invariable levels of standard error may influence the precision of estimation. A CIRT model may allow for the relation between response accuracy and time to be considered in a different way than other assessment measures. A computerized

maze vocabulary task was utilized for this exercise as an alternative to traditional fluency tasks that measure accuracy within a specified time limit and due to the meaningful connection between language skills and comprehension (Scarborough, 2001).

## Applied Example

### *Participants*

A total of 212 third-grade students (110 boys, 100 girls, 2 not recorded) in the southeastern USA participated in the present study. The children came from predominantly low socioeconomic backgrounds, as 70% of the students were eligible for free or reduced price lunch. The sample was primarily White (50%), followed by Hispanic (28%), Black (12%), Asian (4%), Multiracial (4%) and other (2%). Four percent of students were identified as English language learners, and 13% had an individualized education program.

### *Measure*

**Vocabulary Knowledge Task (VKT; Foorman, Petscher, & Bishop, 2012)** In this task, students completed 30 sentences<sup>1</sup> by selecting one of three morphologically related words that best completed the sentence. Items were manipulated to test knowledge of prefixes and derivational suffixes (e.g., The student [attained\*, retained, detained] a high grade in the class through hard work). Because this is a sentence-level task, there are concomitant word recognition, semantic, and syntactic demands in addition to the demands of the phonological and orthographic shifts. Target words in the task were selected on the basis of their printed word frequency (Zeno, Ivens, Millard, & Duvvuri, 1995) and sentences were assigned to grade level using the Flesch–Kincaid grade-level readability formula, along with researchers' judgment about what topics would be familiar to students at different grades. Item administration was such that students were provided a fixed-order set of items and, as each item was presented, students read the sentence, chose the option they believed was correct, and submitted the response via a "submit" button on the screen. Response time was calculated (in seconds) as the amount of time that lapsed from the computer delivery of the item to the student clicking the submit button. Dimensionality was previously evaluated via factor analysis across grades 3–10 (Foorman et al., 2012) and demonstrated that a one-factor model provided the most parsimonious structure to the data. The present study used data from one test form within third grade, allowing for exploration of both the dimensionality of item

---

<sup>1</sup> The 30 sentences administered to students in this example were from an alternate form than that given to a different set of 212 students in the sample from the Petscher et al. (2014) study.

responses within the selected form as well as an item response theory analogue to classical test reliability known as marginal reliability (Sireci, Thissen, & Wainer, 1991).

A benefit of using computerized list or sentence-based item delivery formats is that it may be possible to expand the types of tasks that can include speed as a performance factor. For example, vocabulary knowledge has been measured in multiple formats, including computerized sentence-level maze tasks (Foorman, Petscher, & Bishop, 2012), yet outside of the Test of English as a Foreign Language (Educational Testing Service, 2007), relatively few measures of vocabulary include both accuracy and speed. A student's early vocabulary knowledge is significant predictor of their later reading comprehension performance (Cunningham & Stanovich, 1997; Kamil, 2004; National Institute of Child Health and Human Development, 2000). Understanding the relevance of a student's speed of word knowledge or sentence-level vocabulary comprehension has the potential to increase understanding in this domain.

### *Data Analysis*

Prior to the estimation of the item parameters, the extent to which the items' accuracy responses from the form yielded a unidimensional construct was evaluated. Due to the dichotomous scoring of the item responses, a combination of parametric and nonparametric tests of exploratory and confirmatory factor analyses were conducted in order to comprehensively evaluate the underlying structure of responses (Tate, 2003). Parametric exploratory and confirmatory factor analyses were run using Mplus (Muthen & Muthen, 1998–2012), where the ratio of eigenvalues, comparative fit index (CFI; Bentler, 1990), Tucker-Lewis index (TLI; Bentler and Bonnet, 1980), and the root mean square error of approximation (RMSEA, Browne & Cudeck, 1992) were used to evaluate model fit in the exploratory analysis. All but the ratio of eigenvalues were also used to evaluate the parametric confirmatory factor analysis model. Comparative fit index (CFI) and TLI values greater than or equal to .95 are considered to be minimally sufficient criteria for acceptable model fit, and RMSEA estimates < .05 are desirable. Nonparametric exploratory analysis was run using DIMTEST (Stout, 1987), where a nonsignificant *T* value indicates that the factor structure is essentially unidimensional. DETECT software (Zhang & Stout, 1999) estimated the nonparametric confirmatory model where a DETECT index less than .20 provides evidence of an essentially unidimensional model (Jang & Roussos, 2007).

Following the tests of dimensionality, CTT statistics, including item *p* values, item-to-total correlations, and internal consistency of item responses via Cronbach's alpha, were estimated using SAS 9.3 software (SAS Institute Inc., 2011). IRT analyses using Mplus (Muthen & Muthen, 1998–2012) with maximum likelihood estimation included the fitting of Rasch and two-parameter logistic models in order to estimate item parameters and person-ability scores. The conditional item response theory analyses were fit using the CIRT package (Fox et al., 2007) in R



(R Core Team, 2014). A total of four CIRT models were estimated to identify which best captured the data: (1) a one-parameter response, one-parameter response time model (Model 1), (2) a two-parameter response, one parameter response time model (Model 2), (3) a one-parameter response, two-parameter response time model (Model 3), or (4) a two-parameter response, two-parameter response time model (Model 4). CIRT models were evaluated by using the deviance information criterion (DIC), which estimates the data-model deviation penalized by the model parameters and is computed by the sum of the posterior mean of the deviation (i.e.,  $\bar{D}$ ) and the effective number of model parameters (i.e.,  $p_D$ ). Similar to other information criteria, such as the Bayesian information criterion (BIC), a DIC is evaluated based on its relative comparison to other DICs. As such, while a DIC may be large in magnitude, it is intended to be compared to others, and the model with the smallest DIC should be retained.

### ***Dimensionality Results***

Results from the four methods of testing dimensionality all converged upon the same conclusion, namely, that the item responses were most parsimoniously represented by a unidimensional construct. The analysis of the correlation matrix for the parametric exploratory analysis yielded eigenvalues of 5.52, 1.62, and 1.46 for the first three estimated coefficients. When comparing the ratios amongst them, the ratio of the first to second eigenvalue was 3.39, which was larger than the ratio of the second and third eigenvalues (i.e., 1.12), suggesting that the structure was essentially unidimensional (Divgi, 1980; Lord, 1980). Moreover, the fit for a one-factor solution was excellent, with CFI = .96, TLI = .95, RMSEA = .029 (95% CI = .013, .040). The parametric confirmatory analysis resulted in identical fit indices as the exploratory model. Nonparametric analyses also provided sufficient evidence for a unidimensional structure. A T statistic of  $-.66$  was estimated from the DIMTEST model ( $p = .74$ ), leading to a fail-to-reject decision of the null hypothesis that the item responses were unidimensional in the exploratory model. Similarly, a DETECT index of  $-.0035$  was estimated for the confirmatory model, which was less than the desired  $.20$  for a unidimensional model (Jang & Roussos, 2007).

### ***CTT Results***

Given the evidence for the unidimensionality of the item responses, descriptive statistics for the accuracy of item responses and response times were calculated and reported in Table 7.1; there were no missing data for this sample. The item  $p$ -values ranged from  $.39$  to  $.85$ , indicating a range of difficult to easy items and the average proportion correct was  $.64$ . Internal consistency, as measured by Cronbach's alpha, was initially estimated as  $\alpha = .84$ . Item-to-total correlations were also estimated and

**Table 7.1** Classical and item response properties

Item	<i>p</i> -value	Item-total <i>r</i>	Mean RT	SD RT	1PL		2PL	
					a	b	a	b
1	.75	.24	17.93	12.26	1	-1.23	0.43	-2.61
2	.63	.45	15.62	8.72	1	-0.61	0.99	-0.66
3	.43	.41	19.19	10.75	1	0.29	0.88	0.33
4	.83	.44	15.68	8.93	1	-1.82	1.33	-1.6
5	.64	.48	17.78	10.95	1	-0.68	1.17	-0.66
6	.78	.52	14.11	8.27	1	-1.43	1.71	-1.11
7	.73	.5	20.05	11.92	1	-1.12	1.55	-0.93
8	.53	.47	15.46	11.7	1	-0.17	1.09	-0.18
9	.65	.41	13.89	8.71	1	-0.7	0.93	-0.79
10	.85	.5	13.1	8.41	1	-1.97	2.2	-1.35
11	.45	.29	17.52	11.85	1	0.23	0.55	0.39
12	.67	.26	13.75	7.63	1	-0.8	0.45	-1.61
13	.39	.39	13.31	7.57	1	0.5	0.85	0.58
14	.8	.47	13.36	6.76	1	-1.55	1.41	-1.32
15	.42	.41	14.93	9.72	1	0.38	0.88	0.43
16	.7	.45	16.76	10.84	1	-0.99	1.17	-0.96
17	.69	.58	12.2	6.94	1	-0.94	1.83	-0.73
18	.58	.48	14.75	8.98	1	-0.36	1.16	-0.36
19	.81	.36	13.79	8.96	1	-1.61	0.91	-1.83
20	.6	.6	13.23	6.55	1	-0.48	1.76	-0.4
21	.73	.55	12.76	7.74	1	-1.12	1.78	-0.87
22	.65	.38	15.75	11.53	1	-0.7	0.86	-0.83
23	.65	.32	16.08	10.5	1	-0.7	0.62	-1.07
24	.56	.55	14.54	10.67	1	-0.28	1.48	-0.25
25	.6	.61	12.6	9.68	1	-0.48	1.82	-0.39
26	.52	.48	14.08	9.36	1	-0.12	1.12	-0.13
27	.23	.21	15.72	12.4				

broadly suggested that item responses were moderately associated with overall total test score performance. Item 27 was dropped from subsequent analyses because it was negatively associated the scale. The response-time data indicated that students spent an average of 15.09 s per item ( $SD = 9.46$ ), and ranged from 12.20 s (item 17) to 20.05 s (item 7). Several observations concerning the response-time data are worth noting. The mean and standard deviations were correlated at  $r(1) = .79$ ,  $p < .001$ , which suggested that items on which the students spent the longest also demonstrated the greatest variability in time spent across the sample; however, evaluating the data from Table 7.1 demonstrated that this correlation may vary conditionally on the mean response time. For items where the average response time was long (e.g., items 3 and 7), students tended to vary in their average responses to those items ( $SD = 10.75$  and  $11.92$ ). Conversely, items with short average response times, such as items 19 and 20, presented with standard deviations that illustrate less variability in the average response ( $SD = 8.96$  and  $6.55$ , respectively). Petscher,

Mitchell, and Foorman (2014) systematically show that this variation is significantly more varied for slower average response times versus faster response times.

### *Item Response Theory Results*

For the IRT analyses, Rasch (1PL) and two-parameter logistic (2PL) models were estimated. A comparison of log likelihoods between the two models favored the 2PL model ( $\Delta\chi^2 = 91$ ,  $\Delta df = 26$ ,  $p < .001$ ). Item discrimination and difficulty parameters for both models are reported in Table 7.1. Item difficulties for the 2PL model ranged from  $-2.61$  to  $.58$  and correlated with the classical test  $p$ -values at  $r(1) = -.88$ ,  $p < .001$ . Despite the difference in metrics between the classical and item response approaches, the negative direction of the correlation indicates that items identified as easy in the classical framework (i.e., high  $p$ -value) were also easy in the item response analysis (i.e., negative  $b$  value). The item discriminations in the 2PL model ranged from  $.43$  to  $2.22$ . Similar to the relation between the classical test and item response difficulties, the 2PL discrimination parameter was strongly correlated with the item-to-total statistic at  $r(1) = .91$ ,  $p < .001$ .

### *Conditional Item Response Theory Results*

**Response Time Model Fits** Each of the four CIRT models were estimated and, as part of the model evaluation, it was of interest to evaluate the fit of the models as well as the extent to which resulting theta scores differentially correlated with speed. Scatterplots for the relation between ability and speed are presented in Fig. 7.2. It can be seen that the scatter did not meaningfully differ across Model 1 [ $r(1) = .29$ ,  $p = .003$ ], Model 2 [ $r(1) = .31$ ,  $p = .003$ ], Model 3 [ $r(1) = .29$ ,  $p = .003$ ], or Model 4 [ $r(1) = .32$ ,  $p = .003$ ], and that the relation was moderate in nature such that individuals with higher ability tended to respond to items more quickly. Given the comparability of ability and speed, the model fit was evaluated (Table 7.2). Models 2 and 4 provided the most parsimonious fit as evidenced by the DIC (Model 2 = 12,139, Model 4 = 12,149), whereas Models 1 and 3 were comparatively worse (12,213 and 12,225, respectively).  $\Delta$ DIC values  $> 5$  suggests practically important model fit discrepancies, with the lower value model selected; however,  $\Delta$ DIC = 10 suggests both models should be considered. The  $\Delta$ DIC for Model 2 and 4 compared to Models 1 and 3 was  $> 90$  but the  $\Delta$ DIC between Model 2 and 4 was 10, suggesting that although both Models 2 and 4 were on the threshold of practical difference, they provided superior fit to Models 1 and 3. The primary difference between Models 2 and 4 is the constrained value of the item speed discrimination values to = 1 in the former, and freely estimated values in the latter. Table 7.3 reports the item response parameters (i.e., difficulty and discrimination) and response-time parameters (i.e., intensity and speed discrimination) for Models 2 and 4.

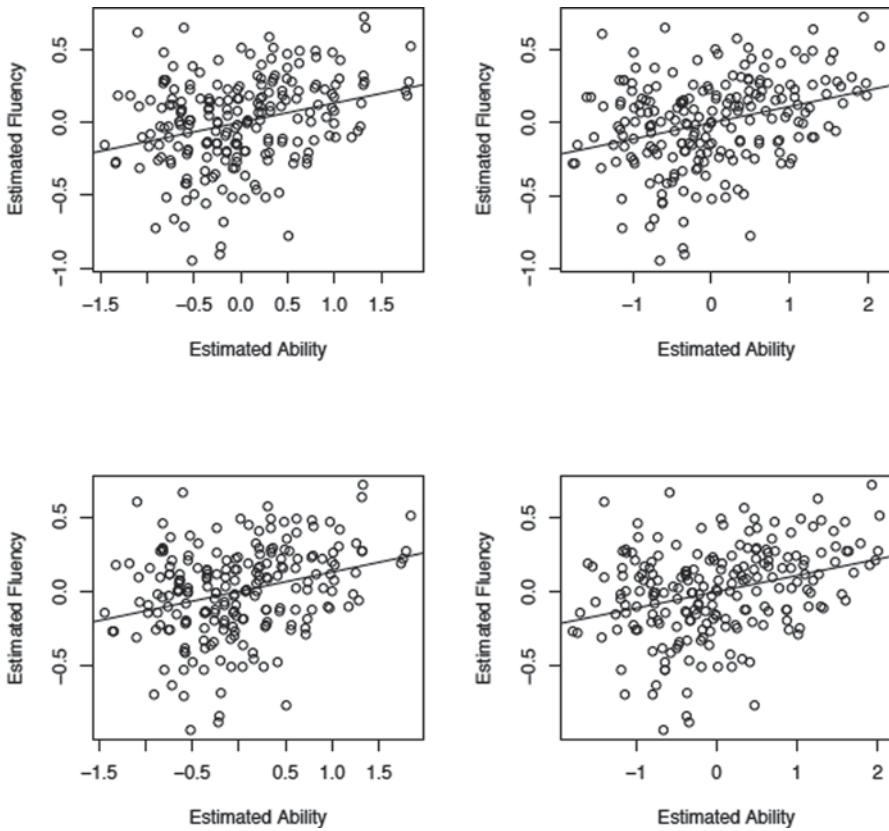


Fig. 7.2 Scatterplots of estimated person ability and fluency scores for CIRT Models 1–4

Table 7.2 CIRT model fit statistics for joint response and response time models

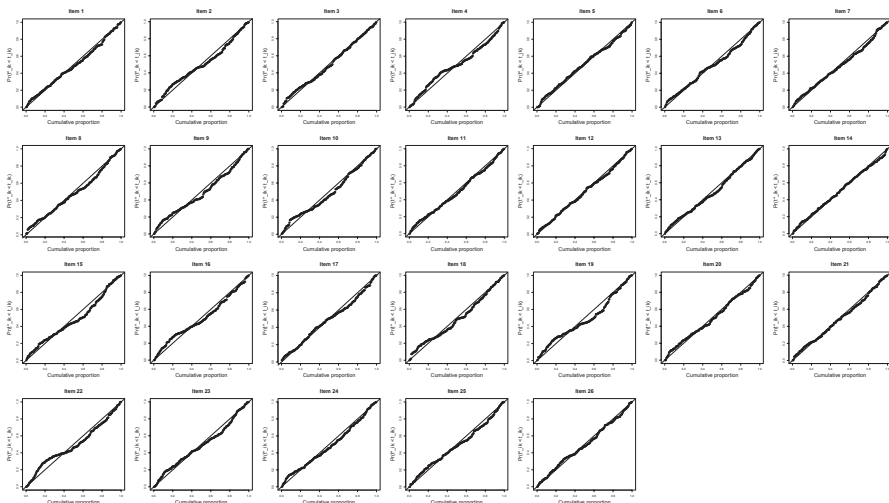
Model	$\bar{D}$	pD	DIC	LL
1	11,748.8	464.0	12,212.8	-5875.2
2	11,665.5	473.6	12,139.1	-5832.2
3	11,737.7	487.3	12,225.0	-5869.5
4	11,652.8	495.7	12,148.5	-5826.5

The results for the item response portion of the models mapped on well to those estimated by the 2PL IRT models. Correlations between the 2PL and CIRT discriminations were near perfect for both Models 2 and 4 [ $r(1)=.98, p < .001$ ] and, additionally, the difficulty values were highly related [ $r(1)=.99, p < .001$ ]. The correlation between the ability score (i.e.,  $\theta$ ) and response speed (i.e.,  $\omega$ ) was  $r(1)=.31$  and  $.32$  (with  $p < .001$ ) for both Models 2 and 4 respectively, which suggested that a moderate association existed between accuracy and speed whereby individuals with higher ability responded to items more quickly than lower ability individuals. In-

**Table 7.3** CIRT response and response time item parameters for models 2 and 4

Item	Model 2				Model 4			
	a	b	$\alpha$	$\beta$	a	b	$\alpha$	$\beta$
1	0.46	-2.49	1.00	2.74	0.46	-2.52	0.78	2.74
2	1.11	-0.60	1.00	2.62	1.10	-0.61	1.05	2.62
3	1.03	0.27	1.00	2.81	1.03	0.28	1.18	2.81
4	1.36	-1.51	1.00	2.63	1.35	-1.52	1.07	2.63
5	1.31	-0.61	1.00	2.73	1.32	-0.61	1.10	2.73
6	1.68	-1.07	1.00	2.51	1.70	-1.07	1.07	2.51
7	1.73	-0.85	1.00	2.85	1.72	-0.85	0.94	2.85
8	1.25	-0.17	1.00	2.56	1.26	-0.18	1.19	2.56
9	1.06	-0.72	1.00	2.49	1.06	-0.72	0.95	2.49
10	2.01	-1.32	1.00	2.42	2.00	-1.32	1.10	2.42
11	0.67	0.32	1.00	2.71	0.68	0.31	1.10	2.71
12	0.51	-1.50	1.00	2.50	0.50	-1.51	0.85	2.50
13	0.95	0.52	1.00	2.46	0.95	0.51	0.88	2.46
14	1.51	-1.22	1.00	2.48	1.50	-1.23	1.05	2.48
15	0.99	0.39	1.00	2.55	0.98	0.39	0.98	2.55
16	1.30	-0.88	1.00	2.68	1.31	-0.88	1.14	2.68
17	2.00	-0.67	1.00	2.37	2.01	-0.67	1.06	2.37
18	1.28	-0.34	1.00	2.56	1.28	-0.34	0.99	2.56
19	0.95	-1.74	1.00	2.46	0.95	-1.75	1.20	2.46
20	1.91	-0.36	1.00	2.48	1.90	-0.36	0.91	2.47
21	1.90	-0.81	1.00	2.41	1.89	-0.81	0.95	2.41
22	0.97	-0.77	1.00	2.60	0.97	-0.77	1.00	2.60
23	0.71	-0.98	1.00	2.62	0.71	-0.98	0.90	2.62
24	1.65	-0.25	1.00	2.49	1.64	-0.25	1.01	2.49
25	1.90	-0.36	1.00	2.34	1.88	-0.36	0.98	2.34
26	1.27	-0.14	1.00	2.48	1.28	-0.13	0.83	2.48

terestingly, the estimated correlations among the item parameters indicated that no relation existed between the item difficulty and intensity [i.e.,  $r(1) = -.02, p = .91$ ]. A review of the estimated intensity parameters (Table 7.3) shows that values do not considerably vary, thus, more difficult items did not require more time. The positive correlation between ability and speed indicates that students with higher abilities spent less time with more difficult questions. We note that it is important to identify that the prior distribution used for response time modeling fits the data being modeled (log-normally distributed). This is evaluated with P-P plots where estimated sample probabilities are plotted against the assumed prior distribution shown as a line in the plot; the extent to which the plotted points deviate from the prior distribution would provide evidence that the sample distribution was biased (i.e., that the sample is not log-normally distributed; Casella & Berger, 2002). Resulting plots (Fig. 7.3) demonstrated that little bias existed when a log-normal distribution was assumed for item response times.

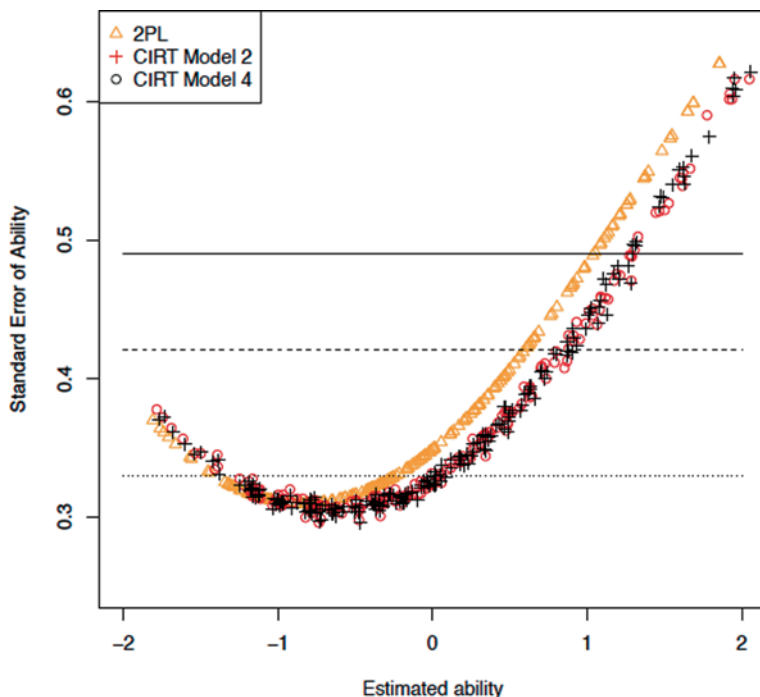


**Fig. 7.3** P-P plots of response time model fits for 26 items in CIRT Model 4 estimation. Plots that lie along the identity line indicate well-fitting response times for the response time model

### *Comparison of Model-Based Reliability*

As previously noted, an expected benefit of the CIRT model is that the standard error of the estimated ability score should be lower when compared to traditional IRT as well as CTT. It is possible to estimate the marginal reliability of scores, with the resulting value allowing for a meaningful comparison to an estimate of internal consistency from CTT (Andrich, 1988; Embretson & Reise, 2000; Orlando & Thissen, 2000). Marginal reliability is computed as a function of the variance of the estimated  $\theta$  scores and the average of the squared standard errors of  $\theta$ . In the present study, the marginal reliability was estimated at .83 for CIRT Models 2 and 4 and .78 for the 2PL model. Compared to the observed reliability in the classical test model ( $\alpha = .81$ ), it is only representative of the average relation between ability and error. A more useful heuristic for evaluating the relation lies in plotting the standard errors for the CIRT and IRT models, as it is possible to view where each model is differentially reliable across the range of abilities as well as how they differ if one were to assume the fixed standard error from CTT. Figure 7.4 plots the standard errors of ability from the 2PL item response model (triangles), CIRT Model 2 (crosses), and CIRT Model 4 (circles). Additionally, three horizontal reference lines are included, which correspond to the observed classical test reliability in the current sample (i.e.,  $\alpha = .81$ ; solid line),  $\alpha = .85$  (dashed line), and  $\alpha = .90$  (dotted line). The standard error associated with each alpha index was converted to an IRT scale in order to allow for a direct comparison between the two theoretical approaches (Dunn, Baguley, & Brunsten, 2014; McDonald, 1999).

Several characteristics of this graph are worth noting; first, both the CTT estimate of internal consistency and the CIRT marginal reliability coefficients assume that the error is constant for all students, as evidenced by the horizontal reference



**Fig. 7.4** Plots of standard error of ability for 2PL IRT Model, CIRT Model 2, and CIRT Model 4. Scale reliability from classical models are included as standard error of measurement for  $\alpha = .81$  (solid line),  $\alpha = .85$  (dashed line), and  $\alpha = .90$  (dotted line)

lines. Based on the plots from item response and conditional item response models, this assumption did not hold. The 2PL item response standard errors varied considerably and were the lowest for individuals whose estimated ability was lower than average ( $\theta < .0$ ). It can be seen that for students whose 2PL IRT ability ranged from  $-2.00$  to approximately  $.50$ , their ability was under the dashed line indicating their scores were minimally reliable at  $\alpha = .85$  and, up to  $\theta$  of about  $.80$ , reliability was equal to the CTT estimate of  $\alpha = .81$ . Conversely, when  $\theta < .80$ , the ability score was less precise, and thus less reliable, than that estimated by CTT. CIRT model reliability was more precise than the 2PL IRT model up to  $\theta$  levels of about  $.80$ . Even within this specified range it can be seen that the standard errors estimated by the CIRT models were below the dotted line, which corresponded to reliability of  $\alpha = .90$ . Toward the upper range of ability scores the advantage of the CIRT models was disappearing.

As the relative difference in standard errors between the IRT and CIRT models varied, conditional on  $\theta$ , it follows that an important contextual consideration is quantifying the conditional impact of the CIRT model on the reliability of resulting  $\theta$  scores for the sample. This was evaluated by computing an estimate of efficiency reflecting the observed percentage change in the standard error of  $\theta$  from the IRT model when response latency was accounted for by the CIRT models. Across the



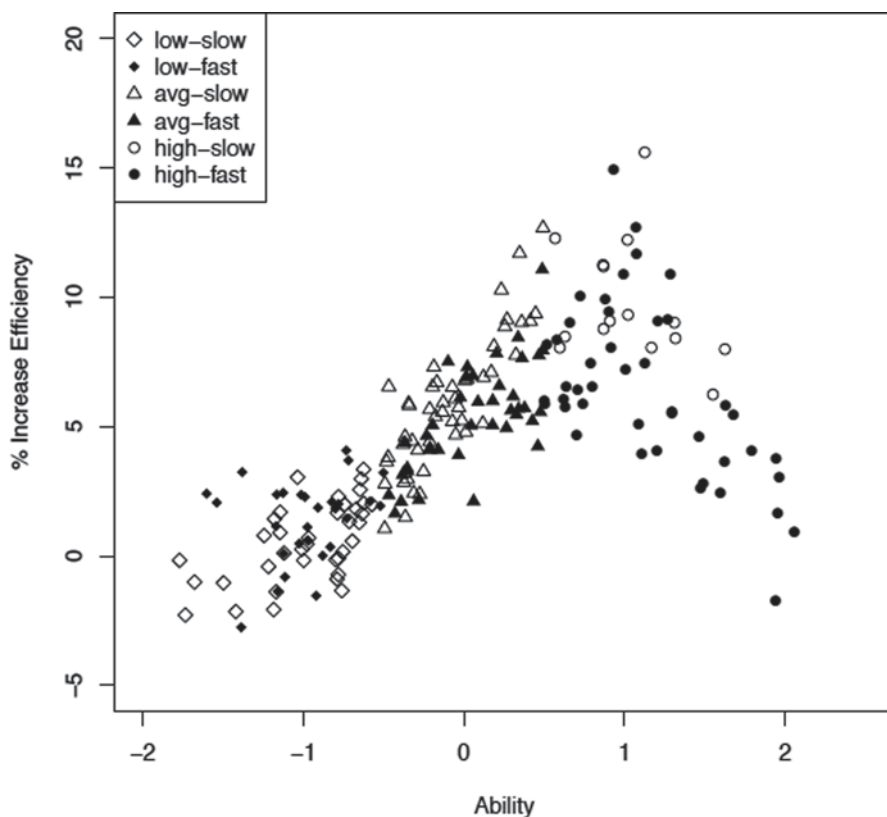
full range of ability scores, the CIRT model (e.g., Model 4), resulted in an average 4.6% reduction ( $SD=3.6\%$ ) in the standard error of individual scores. An analysis of variance was conducted to determine the extent to which efficiency for the CIRT model was greater for low ability ( $\theta < -.50$ ;  $N=64$ ), average ability ( $-.50 < \theta < .50$ ;  $N=90$ ), or high ability ( $\theta > .50$ ;  $N=58$ ) individuals in the sample. Results indicated a strong effect for ability groups in efficiency [ $F(2209)=100.9$ ,  $p < .001$ ], with students who were categorized as low ability gaining little from the CIRT model ( $M = .93\%$ ,  $SD = 1.60\%$ ) compared to either the average ( $M = 5.61\%$ ,  $SD = 2.32\%$ ) or high ability ( $M = 7.23\%$ ,  $SD = 3.39\%$ ) students. All pairwise comparisons were statistically significant ( $p < .001$ ), with Hedge's  $g$  effect sizes demonstrating that efficiency of the model was stronger for average ability compared to low ability students ( $g = 2.27$ ), as well as high ability compared to either low ( $g = 2.40$ ), or average ( $g = .58$ ) ability differences.

Such stark differences in the model efficiency in favor of the high-ability students, coupled with higher variance in efficiency for those individuals, warranted further exploration. A scatter plot was generated (Fig. 7.5) that plotted ability against the CIRT model efficiency for the full sample. Within this plot the three ability groups are denoted by different markers and they are further distinguished by whether the student was below the mean in average response time (i.e., faster) or above the mean (i.e., slower); fast students are represented by the open shapes and slower students by the filled shapes. It can be observed that low- and average-ability students have an approximately equal number of individuals who were fast or slow, whereas the high-ability students maintained a stronger representation of fast students. The nonlinear shape of the scatter is largely marked by the variation of these high-ability, fast-response students, in that some of these individuals received a precision benefit to their ability score, while others did not.

To more fully explore this phenomenon, elements of Figs. 7.4 and 7.5 were combined to evaluate the relations among ability, the standard error of ability, and the percent efficiency for the CIRT model (Model 4; Fig. 7.6). This plot highlights what has been previously observed, namely, that CIRT model efficiency is strongest for the higher ability students, as well as that the standard errors for the full sample are lowest for lower ability individuals in the sample. What this further illuminated was that there appeared to be diminishing returns in response speed as it pertained to ability and its standard error. Note that as ability and standard error increased, so did the impact of the efficiency of the CIRT model, yet once  $\theta$  exceeds 1, model efficiency showed a decline in returns.

## Conclusions

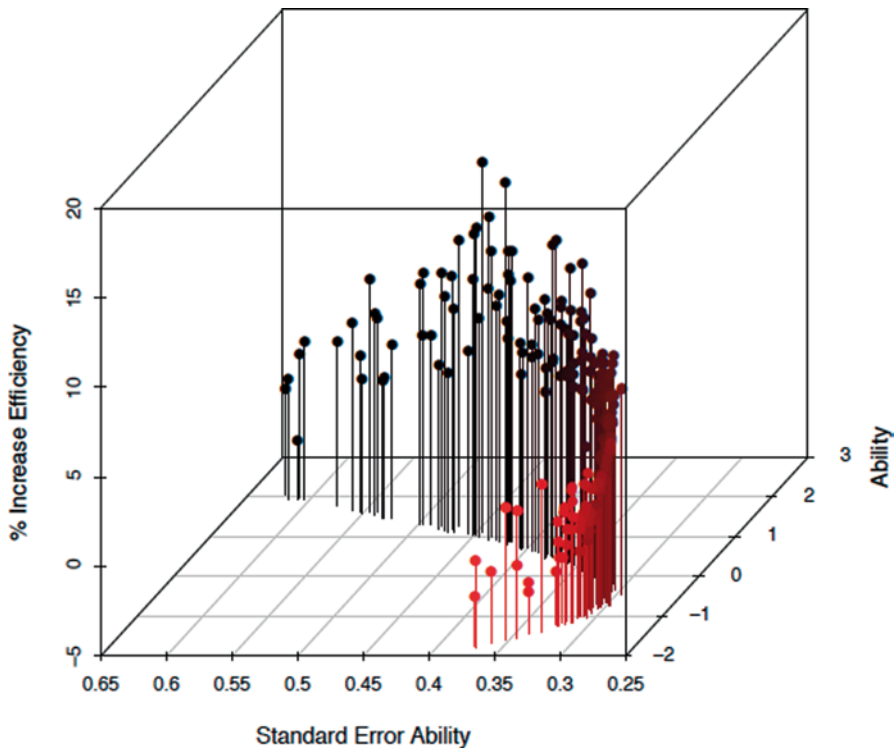
Determining comprehensive and efficient ways to measure student performance is essential for individuals working in school settings as well as those conducting applied research. The current chapter considered the effects of incorporating response time into an item response measurement framework. First, in order to consider the



**Fig. 7.5** Plot of estimated ability score and CIRT model efficiency gain, with scores grouped by score and response time. Low, medium, and high scores, and fast and slow responses provide grouping characteristics

general impact on this process in terms of reliability, the estimated item parameters from CTT, IRT, and CIRT analyses were compared. Relations between response accuracy and speed were specifically studied by evaluating the extent to which the three measurement approaches provided different information concerning the reliability or precision of student ability scores. Through this comparison, the added value of including response time above the information provided by accuracy alone was considered. Item parameter values were very high across the three approaches. CTT  $p$ -values were strongly associated with IRT and CIRT item difficulties, item discriminations, and item-to-total correlations. Further, item response analysis of the data found variability in ability scores connected to where on in the distribution the ability score was estimated. This suggests a benefit to IRT estimations compared to static CTT estimations.

The CIRT model improved on the reliability of student scores by yielding lower overall standard errors associated with the individual ability scores. But the extent to which the reliability improved was contingent on the specific ability level of the individual. This finding is notable because speed may be a less significant factor in



**Fig. 7.6** 3D scatterplot of standard error of ability, ability, and percent efficiency gain in CIRT modeling. Increasing score values fade to black, with higher efficiency gains indicated by more elevated points

an instance when one's ability on the administered task is high and the precision of the ability score is low. Regardless of this contingency, these results indicated that response time is a valuable consideration and can be incorporated into a measurement model. Recording response times is an easily built-in component of computerized testing and, thus, researchers do not need to expand testing time in order to capture data that could be used for modeling purposes. Rather, the response time simply needs to be recorded by the software program and recovered in a data file along with the accuracy of the student response. Further, it is relevant to note that the CIRT analysis is no more difficult to conduct than is a traditional IRT analysis, and may, in fact be easier due to the specificity of the software available to conduct the analysis (R package version 3.0.0; R Core Team, 2014).

Use of a vocabulary knowledge task for this analysis extends the work that has previously been done in terms of evaluating CIRT models. This application may be used to consider both accuracy and speed in estimating performance in a number of areas and has meaningful connections to the work of several groups. Educators in practical settings are continuously looking for efficient and reliable testing methods, while researchers are invested in improving the precision of assessment scores. CIRT models may lend themselves to using response time to meet these

goals. By understanding not only what students know, but also how facile they are with this knowledge, it may be possible to gather greater understanding that can connect to performance outcomes.

## Considerations for Future Work

Students in this illustration were not aware that item-level response time was being collected and, thus, it is possible that differences in precision could be obtained when students are aware that their rate of response is being considered as a performance factor. If accuracy and response-time data are available at the item level, future work might test the dimensionality of both components to see if evidence is presented for a unidimensional or multidimensional representation of the data. With the implementation of computers in task assessment, the introduction of adaptive methods in testing is also productive. The added benefit of improved reliability when response time data is used is that researchers may use the CIRT model to create an adaptive task that implements both response and response time data, and balance reliability with number of items presented (Prindle, 2012). The balance of items necessary is task dependent on the relation between individual accuracy and speed.

## References

- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.
- Bentler, P. M., & Bonnet, D. C. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.
- Blackwell, C. K., Lauricella, A. R., Wartella, E., Robb, M., & Schomburg, R. (2013). Adoption and use of technology in early education: the interplay of extrinsic barriers and teacher attitudes. *Computers & Education*, *69*, 310–319. doi:10.1016/j.compedu.2013.07.024.
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, *21*, 230–258.
- Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Pacific Grove, CA: Duxbury.
- Cattell, R. B. (1948). Concepts and methods in the measurement of group syntality. *Psychological Review*, *55*, 48–63. doi:10.1037/h0055921.
- Chard, D. J., Vaughn, S., & Tyler, B. (2002). A synthesis of research on effective interventions for building reading fluency with elementary students with learning disabilities. *Journal of Learning Disabilities*, *35*(5), 386–406. <http://search.proquest.com/docview/619935634?accountid=4840>.
- Christ, T. J., & Silbergitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review*, *36*, 130–146.
- Cummings, K. D., Atkins, T., Allison, R., & Cole, C. (2008). Response to Intervention. *Teaching Exceptional Children*, *40*, 24–31.
- Cummings, K. D., Park, Y., & Schaper, H. A. B. (2012). Form effects on DIBELS Next oral reading fluency progress-monitoring passages. *Assessment for Effective Intervention*, *38*, 91–104.
- Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*(6), 934–945.

- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford.
- Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184–192.
- Divgi, D. R. (1980). Dimensionality of binary items: Use of a mixed model. Paper presented at the annual meeting of the National Council on Measurement in Education. Boston.
- Dunn, T. J., Baguley, T., & Brunsten, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105* (3), 399–412.
- Educational Testing Service. (2007). *Test and score data summary for TOEFL internet-based test*. Princeton: Author.
- Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah: Erlbaum.
- Ferrando, P., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement, 31*, 525–543. doi:10.1177/0146621606295197.
- Foorman, B. R., Petscher, Y., & Bishop, M. D. (2012). The incremental variance of morphological knowledge to reading comprehension in grades 3–10 beyond prior reading comprehension, spelling, and text reading efficiency. *Learning and Individual Differences, 22*, 792–798. doi:10.1016/j.lindif.2012.07.009.
- Fox, J. P., Klein Entink, R. H. K., & van der Linden, W. J. (2007). Modeling of responses and response time with the package CIRT. *Journal of Statistical Software, 20*, 1–14.
- Francis, D. J., Santi, K. S., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*, 315–342. doi:10.1016/j.jsp.2007.06.003.
- Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21*(2), 449–460.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: a theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239–256. doi:10.1207/S1532799XSSR0503\_3.
- Goodglass, H., Theurkauf, J. C., & Wingfield, A. (1984). Naming latencies as evidence for two modes for lexical retrieval. *Applied Psycholinguistics, 5*, 135–14.
- Gray, L., Thomas, N., & Lewis, L. (2010). Teachers' use of educational technology in U.S. public schools: 2009 (NCES 2010-040). Retrieved from the U.S. Department of Education, National Center for Educational Statistics, Institute of Education Sciences. <http://nces.ed.gov/pubs2010/2010040.pdf>.
- Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement, 44*, 1–21.
- Kamil, M. L. (2004). Vocabulary and comprehension instruction: Summary and implications of the national reading panel findings. In P. McCardle & V. Chhabra (Eds.), *The voice of evidence in reading research* (pp. 213–234). Baltimore: Paul H Brookes.
- Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., and Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods, 14*, 54–75. doi:10.1037/a0014877.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York: Erlbaum Associates.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah: Lawrence Erlbaum Associates, Inc.
- Mercer, S. H., Dufrene, B. A., Zoder-Martell, K., Harpole, L. L., Mitchell, R. R., & Blaze, J. T. (2012). Generalizability theory analysis of CBM maze reliability in third- through fifth- grade students. *Assessment for Effective Intervention, 37*, 183–190. doi:10.1177/1534508411430319.
- Miranda, H., & Russell, M. (2011). Predictors of teacher-directed student use of technology in elementary classrooms: A multilevel SEM approach using data from the USEIT study. *Journal of Research on Technology in Education, 43*, 301–323.
- Muthen, L. K., & Muthen, B. O. (1998–2012). *Mplus. Seventh Edition*. Los Angeles: Muthen & Muthen.

- National Institute of Child Health and Human Development. (2000). Report of the National Reading Panel. Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4769).
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64.
- Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology*, *67*, 461–469.
- Petscher, Y., & Kim, Y. S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology*, *49*, 107–129. doi:10.1016/j.jsp.2010.09.004.
- Petscher, Y., Mitchell, A. M., & Foorman, B. R. (2015). Improving the reliability of student scores from speeded assessments: an illustration of conditional item response theory using a computer-administered measure of vocabulary. *Reading and Writing*, 1–26.
- Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment*, *23*, 326–338. doi:10.1177/073428290502300403.
- Pressey, B. (2013). Comparative analysis of national teacher surveys. [http://www.joanganzcooney-center.org/wp-content/uploads/2013/10/jgcc\\_teacher\\_survey\\_analysis\\_final.pdf](http://www.joanganzcooney-center.org/wp-content/uploads/2013/10/jgcc_teacher_survey_analysis_final.pdf).
- Prindle, J. J. (2012). A functional use of response time data in cognitive assessment. (Doctoral dissertation). Retrieved from USC Digital Library.
- R Core Team. (2014). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- SAS Institute Inc. (2011). Base SAS® 9.3 Procedures Guide. Cary: SAS Institute.
- Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. Neumann & D. Dickinson (Eds.), *Handbook for research in early literacy* (pp. 97–110). New York: Guilford.
- Scheiblechner, H. (1985). Psychometric models for speed-test construction: The linear exponential model. In S. E. Embreston (Ed.), *Test design developments in psychology and psychometrics* (pp. 219–244). New York: Academic Press.
- Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: building the foundation for future assessments*. Mahwah: Lawrence Erlbaum Associates.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247. doi:10.1111/j.1745-3984.1991.tb00356.x.
- Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421–457.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, *52*, 589–617.
- Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*, 159–203.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. doi:10.1007/s11336-006-1478-z.
- van der Linden, W. J. (2011). Modeling response times with latent variables: principles and applications. *Psychological Test and Assessment Modeling*, *53*, 334–358.
- van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect responses in computerized adaptive testing. *Psychometrika*, *68*, 251–265.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.
- Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. New York: Touchstone Applied Science Associates, Inc.
- Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

## Chapter 8

# An Introduction to the Statistical Evaluation of Fluency Measures with Signal Detection Theory

Keith Smolkowski, Kelli D. Cummings and Lisa Strycker

Fluency represents the learned ability to respond quickly, effortlessly, and accurately to a given stimuli. The basic principle applies to skills across a variety of domains, from sports to science, music to mathematics, and reading to public speaking. Athletes often refer to fluency in terms such as muscle memory or motor learning, forms of procedural memory. Educators may refer to fluency as automaticity or overlearning. Regardless of the specific terms, fluency captures the ability to perform a task correctly and quickly without conscious, focused thought about the details. Hence, fluency comprises two key features of a skill: speed and accuracy.

Fluency can be described on a continuum. Many small children begin to learn counting around 2 or 3 years of age, and these early learners count slowly and often only to three or maybe five or ten. They also make errors, possibly skipping a number or getting two out of order. As children age, they learn to count more quickly and more precisely, with most able to count into the 20s by age 5 and to 50 by the end of kindergarten (Clarke, Baker, Smolkowski, & Chard, 2008). Older children and adults can count almost indefinitely and without much thought; it becomes automatic. The same basic continuum of fluency applies to learning a variety of activities such as singing a song, playing chess, reading a book, dancing, tying a shoelace, or driving a car. Whether someone is sufficiently fluent at a skill depends on context. Reading at 40 words/min would be outstanding for a kindergarten student but far short of adequate for a third grader.

Fluency in many skills also requires fluency with subskills. To become fluent at reproducing a piece of music, a cellist must become skilled at reading music,

---

K. Smolkowski (✉) · L. Strycker  
Oregon Research Institute, Eugene, Oregon, USA  
e-mail: keiths@ori.org

L. Strycker  
e-mail: lisas@ori.org

K. D. Cummings  
University of Maryland, College Park, USA  
e-mail: kelic@umd.edu



intonation, complex rhythms, tempo changes, vibrato, and other techniques (e.g., pizzicato and bowing). Instruction does not begin with whole pieces of music, but specific subskills. A student might first learn to pull her bow across the strings, and after learning how to reduce the screech to a pleasing tone, she might begin to learn finger placements, then rhythms, and so on. Similarly, dancers perform best through *marking*, where they practice only partial movements of a performance, that conserves energy and reduces cognitive load (Warburton, Wilson, Lynch, & Cuykendall, 2013). No individual skill, mastered to perfection, can by itself approach the beauty achieved by masters such as Yo-Yo Ma. Automaticity or fluency at each skill, and each subskill, requires practice (Bengtsson et al., 2005; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Ericsson, Roring, & Nandagopal, 2007; Fields 2005; Logan 1988; Posner, DiGirolamo, & Fernandez-Duque, 1997). “The idea that practice can automate a skill has been with us since the inception of psychology” (Posner et al., 1997, p. 267). Indeed, fluency requires deliberate practice, which includes repeated, independent activities designed to focus on specific skill deficits, account for preexisting knowledge, and provide immediate feedback (Engelmann & Carnine, 1991; Ericsson, Krampe, & Tesch-Römer, 1993; Kopiez & Lee, 2006; Smolkowski & Gunn, 2012).

When learning to read, children must decode the sounds of specific letters, blend them together into a word, and produce a coherent thought from several successive words. A kindergarten student who slowly sounds out letters and then carefully figures out the words in a short sentence would be considered to be *disfluent* or low on the fluency continuum. Adults, many of whom long forgot about their struggles learning to read, simply see words and phrases, as if the brain “magically” understands them. This behavior represents the other end of the spectrum—fluency. Reading requires fluency with several skills. As English, like many other languages, is code based, a reader must learn that words are formed from individual sounds, that written letters codify those sounds, that blending those sounds together forms whole words, and that reading a string of words produces a coherent thought. A student who cannot produce all his consonant sounds or associates a particular sound (e.g., /d/) with the wrong letter (e.g., b) will not likely become fluent at reading whole words unless intervening instruction is provided. Similarly, students who become fluent at reading words but fail to understand the text have not likely become fluent with their vocabulary or comprehension skills. Such problems usually represent instructional errors—a focus on a subset of component skills (e.g., letter sounds and blending) while paying insufficient attention to others (e.g., word meaning and comprehension). The fluent application of a skill requires frequent practice on all relevant subskills (Ericsson et al., 2007). Fluency with decoding does not necessarily imply fluency with comprehension or vice versa.

Fluency with a skill requires deliberate practice, and not simply repetition of subskills that have already been grasped, but rehearsal of those subskills that have not yet become fully fluent. A cellist practicing for a production of Bach’s Brandenburg Concerto No. 3 would not spend her time on the portions of the piece that she found easy, but rather she would focus on the most challenging sections, such as those

with complicated rhythms and transitions and the coordination with the violins and violas. Similarly, a student who supposedly “word calls” but cannot comprehend (cf. Hamilton & Shinn, 2003) has likely over-practiced decoding with insufficient practice on comprehension skills. This might be similar to the cellist whose focus on intonation precludes her attention to the rhythms and tempo changes that produce the “feel” of a piece. Practice on subskills that students have mastered can become easy and rewarding, so students often need help identifying and spending the time on the areas that require additional rehearsal.

When students do not reach fluency, diagnosing the source of the disfluency becomes critical for educators, conductors, and coaches. Educators often assess fluency with measures that indicate the rate of accurate responding, such as a measure of oral reading fluency (ORF) that characterizes fluency with the number of correct words read in 1 min. As schools, by design, begin with students who are not yet fluent at the relevant skills, educators need to know how to best identify the strengths and weaknesses of their students. It is important to identify the students who struggle with fluency, ideally early enough that teachers can provide additional supports before their students fall too far behind. The answer to the following question helps teachers discriminate students’ challenges and offer targeted supports: What skills can children perform with speed and accuracy and with which skills do they struggle? That is, if students struggle, do they have difficulty with all subskills or only a subset?

This approach, however, requires a judgment about the level at which poor fluency becomes disfluency, which also depends on the context. With guidance on acceptable levels of fluency performance in different grade levels and different times of the year, teachers, school psychologists, or other educators can make informed instructional decisions about the provision of services for students. In many such cases, educators must make decisions based only on their own familiarity with their students and their knowledge of the instructional material; they make professional judgments. Judgments, however, are not without problems. Teachers, like all people, fall prey to a number of cognitive heuristics and biases (Connolly, Arkes, & Hammond, 2000; Kahneman, Slovic, & Tversky, 1982) that reduce the accuracy of instructional decisions. The *halo effect* (Dompnier, Pansu, & Bressoux, 2006), for example, refers to the case where judgments about specific skills become influenced by overall impressions (e.g., friendliness, appearance, or skill in another domain). Rather than relying solely on judgments, teachers can more easily and accurately assign students to small-group instructions or other support services through the use of *diagnostic* or *classification* decisions methods. This chapter describes the basic methods recommended for the development and evaluation of classification systems using a framework called signal detection theory. The methods can be applied to any screener or test (continuous or ordinal) used to gauge the likely accomplishment of some relevant criterion, including many measures that are available in education.

## Statistical Evaluation of Screeners

The methods for the statistical evaluation of screeners attempt to first characterize the overall accuracy of a continuous or ordinal measure to determine if it can adequately discriminate those who meet the criterion versus those who do not and then to identify the score that optimally predicts membership in one of two populations. The methods answer questions like the following: Can a particular measure of fluency discriminate between a population of normally achieving students and a population of students with identified learning disabilities, and what precise level of fluency best discriminates between the two populations? The two populations, however, could be represented by any classification, including artificial populations of students who “succeed” or “fail” on a particular criterion test. To catch students who struggle with mathematics concepts, for example, their teachers would benefit from a universal screening system that allows for early detection of difficulties, before the students have the opportunity to develop more serious performance problems. A well-developed diagnostic system will suggest the level at which a student is sufficiently likely to encounter problems later so that the teacher can then choose the appropriate level of supports to prevent future struggles.

A screening system is one representation of a diagnostic or classification system that aims to classify individuals into one of two categories based on a screener. The *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS) along with the information used to identify levels of risk for individual students represents one screening system. A single DIBELS measure, by itself, is just a screener, and it becomes a diagnostic system when coupled with the guidance about how to classify students into risk categories based on their scores. Decisions about which interventions students may require lie outside the screening system, as do data management services (e.g., AIMSweb or the DIBELS Data System). Data management services provide easy access to data entry tools and help teachers classify their students based on cut scores, but do not represent screening systems per se. The DIBELS Data System, for example, offers educators access to both the DIBELS and easyCBM screening systems, but teachers have the option to use either screening system independent of the DIBELS Data System. The underlying set of screening measures and decision rules represents the screening system.

Diagnostic systems are intended to help make dichotomous decisions and include different types of tests, such as *screeners* and *diagnostic tests*. Diagnostic tests typically classify individuals with signs or symptoms of a disability or disorder. Screeners are typically given to a larger sample where such signs are not yet present. If a student was assigned to Tier 3 based on the results of a screener, a special educator might then use a diagnostic test to determine if the child might be diagnosed with a learning disability. Although educators and education researchers frequently refer to academic screeners to aid decision-making, most such tests are technically *prognostic tests* (Kraemer, 1992). Screeners and diagnostic tests assume that the disorder or disability is present during the period of testing, while prognostic tests predict a disorder or disability in a follow-up period. As research on fluency tests generally

use a criterion collected in a follow-up period, well after the screener administration and additional instruction, the prognostic-test label better describes this class of assessments. Nonetheless, all three types of tests are intended to discriminate two categories of individuals and have been treated identically with respect to the evaluation methods (Kraemer, 1992; Pepe, 2003), so we continue the tradition in education and refer to fluency screeners rather than fluency prognostic tests.

The goal of diagnostic or classification systems is to demonstrate the accuracy of a screener or diagnostic test, and to choose a cut score or *decision threshold* among the many possible scores along the range of the diagnostic measure that best corresponds to some important student outcome. The methods attempt to do so in an unbiased and consistent manner, so all students are classified identically, within the limits of measurement error. To align with the variety of supports offered by schools, a tiered set of decision thresholds can also be useful. Each decision threshold can allow educators to make well-informed decisions that may lead to the provision of additional supports, such as increasing practice opportunities, placing students in smaller instructional groups, or referring them for special education.<sup>1</sup> The value of screening systems depends on how well they meet certain methodological standards and how well teachers and their schools implement the system (Cook & Odom, 2013).

Diagnostic systems have been applied across a variety of settings. “Diagnostic systems are all around us. They are used to reveal diseases in people, malfunctions in nuclear power plants, flaws in manufactured products, threatening activities of foreign enemies, collision courses of aircraft, and entries of burglars” (Swets, 1988, p. 1285). The methods were first developed to work on radio signal detection (Peterson, Birdsall, & Fox, 1954), and this history accounts for some of the unusual languages, such as “signal detection” or “receiver operating characteristics” in the literature. The methods, however, have successfully spread to medicine (Kraemer, 1992; Pepe, 2003; Zhou, McClish, & Obuchowski, 2002), epidemiology and public health (Fleiss, 1981; Katz & Foxman, 1993), psychology (Swets, 1996; Swets, Dawes, & Monahan, 2000b), weather and forecasting (Brooks, 2004; Mason & Graham, 1999), and elsewhere (e.g., Burkel et al., 2002).

The work on signal detection has also benefited from Paul Meehl’s (1954) book, *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, and his criticism of prior attempts to create cut scores for classification systems (Meehl & Rosen, 1955). Meehl (1954) first showed that statistical models outperformed clinical judgment when making predictions about treatments, which has survived the test of time unscathed (Grove, 2005; Grove & Lloyd, 2006; Grove et al., 2000; Meehl, 1986). Meehl’s work has since been extended to show that regression weights from statistical models are, in fact, unnecessary and that

---

<sup>1</sup> Screening systems may ultimately be used to discriminate between students with a specific disability and typically achieving students, but the identification of a specific disability requires more information and is a more-involved process than simply labeling all students who fall below a certain screener score (AERA, APA, & NCME, 1999; Engelmann & Carnine, 1991; Smolkowski & Cummings, 2014). We therefore refer generically to students with “difficulties” in reading, math, or other content rather than students with disabilities.

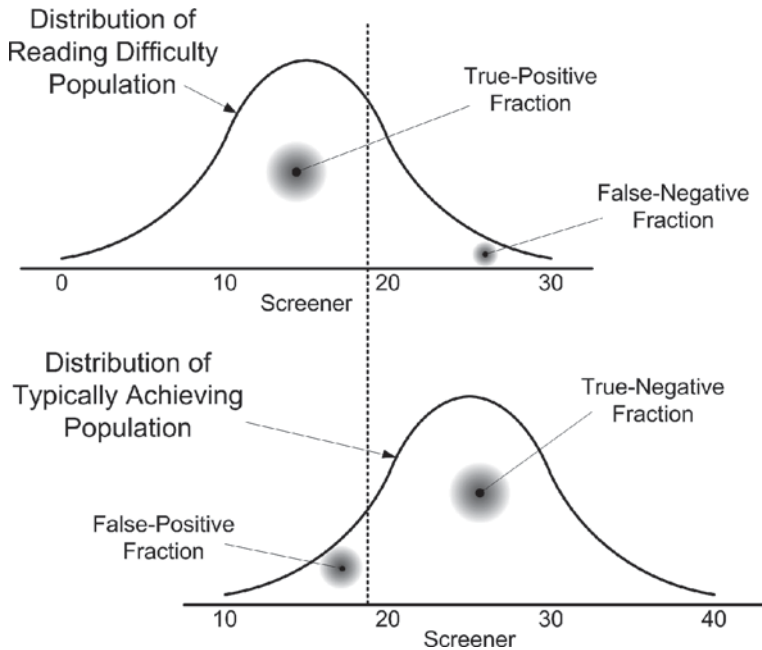
correlation weights (Goldberg, 1972) and even unit weights (Dawes, 1979, 1986) can, in certain situations (Dana & Dawes, 2004), outperform both clinical and even complex statistical models. Swets et al. (2000b) showed how this work extends to screening systems. They demonstrated that the improvements of statistical prediction are most beneficial for dichotomous-choice decisions when cut scores have been chosen through a set of rigorous methods. The approach to diagnostic decision-making, then, has a rich history in the behavioral sciences, and offers a number of valuable tools for education researchers.

In this chapter, we discuss the accuracy of screeners used in diagnostic systems and the theoretical foundations of diagnostic methods. We then move on to decision criteria—the choice of a cut score—as well as the various indices and rules. Before proceeding, we make a brief note about terminology. Educators and educational researchers prefer to frame topics positively, in terms of success. The literature on diagnostic decision systems, in contrast, typically frames the methods and discussions around adverse events. These events include the prediction of tornados, the presence of cancer, fractures in critical airplane components, and mental disorders. Consequently, in most papers and books written about methods for diagnostic systems, a higher score indicates an adverse event. We refer to student with reading difficulties or failure, or risk of reading difficulty or failure, where higher scores on reading screeners indicate better performance and lower scores indicate difficulties. To reduce confusion, we have reversed the scaling direction in the examples presented in this chapter to match common examples in education, where higher scores are better.

## Diagnostic systems

The theoretical representation of a diagnostic or classification system includes two populations, a diagnostic test, and four potential outcomes. The two populations are represented by two distributions, one for the typically achieving student population and one for the population of students with reading difficulties. The two population distributions are frequently determined by the criterion measure—a “gold standard” test—but in general, students can be assigned to the two populations based on any method that validly determines population membership. Criterion measures are also generally imperfect (see below), but signal detection theory nonetheless assumes that the goal is to discriminate between two distinct populations.

The two distributions can also be characterized by the scores from a diagnostic test or screener. The exemplar diagnostic system depicted in Fig. 8.1 attempts to discriminate between two populations, a reading difficulty population and a typically achieving population, based on a screener. The population distributions are characterized by scores on the screener with values that range from 0 to 30 for the reading difficulty population and from 10 to 40 for the typically achieving population. A lower score on this screener indicates a greater chance that the student is a member of the reading difficulty population. The two populations overlap in terms of screener scores, and a different screener that more accurately classified students would lead to less overlap among the distributions—the distribution of the typically



**Fig. 8.1** Overlapping populations for typically achieving students and students with reading difficulties. The vertical dashed line defines the decision threshold used to discriminate between population memberships, where a positive test indicates membership in the reading difficulty population. The decision threshold determines the fractions of students classified into four possible outcomes: the *true-positive fraction* (TPF) and *false-positive fraction* (FPF) associated with students in the typically achieving population and the FPF and TPF associated with the reading difficulty population

achieving students in Fig. 8.1 would shift to the right. A good system would allow an educator to make a relatively accurate dichotomous decision about whether to assume that a given student is a member of the typically achieving population or the population of students with reading disabilities.

A sufficiently low score on the screener indicates a *positive* test result or a decision that the student is a member of the reading difficulty population and a *negative* test result, a sufficiently high score, implies that the student is a member of the typically achieving population. In Fig. 8.1, the vertical dashed line depicts the decision threshold, the cut score or cut point along the screener used to classify students into one of the two distributions. Scores to the left of the line are considered to be a positive test result and scores to the right indicate a negative test result.

Figure 8.1 also shows the four possible outcomes of a dichotomous screening decision. A negative decision—a score to the right of the dashed line—may be a *true-negative*, if the student is really in the typically achieving population, or a *false-negative*, if the student is truly in the population of students with reading disabilities. Similarly, a positive test result occurs when scores fall the left of the dashed line and may represent a *false-positive*, if the student is truly in the typically achieving population or a *true-positive* if the student is in the reading difficulty population.



A high-quality diagnostic system will maximally discriminate between the two populations. Hence, the methods developed to evaluate the system represented in Fig. 8.1 must address two concerns: the accuracy of the screener—or equivalently the discrimination between the two populations—and the choice of the decision threshold.

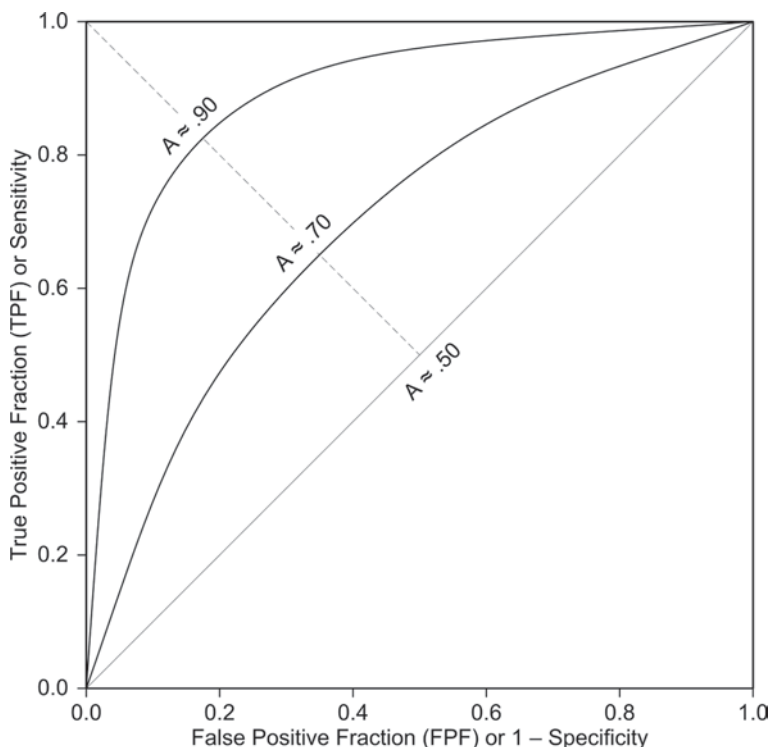
## Accuracy of Diagnostic Systems

This section discusses the accuracy of a diagnostic system and primarily concerns the overall accuracy of a screener to determine if it sufficiently describes the underlying populations of interest or to determine how it compares with other screeners regardless of a chosen decision threshold. Although selecting a decision threshold may be the ultimate goal, in many cases a different decision threshold could be used in different situations. Furthermore, comparing two screening systems by using, say, the sensitivity of a single cut point from each leaves out considerable information that could become valuable in later research. When comparing overall screener accuracy, irrespective of an a priori decision threshold, it is important to recognize that the methods we outline below assume *the same general population of learners*. Screener characteristics, and the performance of a criterion measure, may differ between populations of children from the Spanish-speaking homes of recent immigrant families or children with hearing deficits. Nonetheless, a screener determined to be superior with one population may prove similarly more accurate with different populations.

A more accurate screener will better discriminate between the two underlying populations of interest (e.g., students with reading difficulty and those without)—it will result in fewer false-positives and false-negatives—than a less accurate screener. In terms of Fig. 8.1, a more accurate screener will result in greater separation between the two population distributions. To evaluate whether or not a diagnostic test is accurate, we recommend three evaluative criteria (Pepe, 2003; Swets, 1988, 1996). First, a measure of accuracy must be independent of the probability of occurrence of the criterion event of interest (e.g., the proportion of students with reading difficulties). In the present case, this means that the estimate of the accuracy of a particular measure should not depend on whether 15% of the sample or 50% of the sample has failed to achieve the desired reading outcomes, provided the samples represent the same overall population of students. Second, the system's accuracy should remain unaffected by the criterion measure. This means that estimates of accuracy should not rely on the diagnostic system under consideration, so a system should not be tested against itself; it should be evaluated against an external criterion. Finally, the accuracy of a screener should not depend on the specific decision threshold (i.e., cut score) used to predict membership in the two populations. In this section we demonstrate the methods used to achieve these fundamental attributes in the assessment of screener accuracy.

The receiver (or relative) operating characteristic (ROC) curve has become the standard for the evaluation of accuracy, and the area under the curve,  $A$ , is the recommended index of accuracy (Swets, 1996; Pepe, 2003). An ROC graph represents the proportion of times that an adverse outcome was correctly chosen





**Fig. 8.2** The relative operating characteristic (ROC) graph plots the true-positive fraction (sensitivity) against the false-positive fraction (1 - specificity). Of the two curves depicted, one represents a fairly accurate screener with the areas under the curve,  $A \approx .90$ , and the other a less accurate screener, with  $A \approx .70$ . The diagonal line would represent a screener that carries no information

by the screener relative to the proportion of times that an adverse outcome was incorrectly chosen across the range of all possible screener scores. The first proportion concerns true-positives or “hits” and the second involves false-positives or “misses.” The ROC curve is a plot of the true-positive fraction (TPF) on the vertical axis against the false-positive fraction (FPF) on the horizontal axis as the decision threshold changes from the lowest score on the screener to the highest, that is, as the vertical line in Fig. 8.1 moves from left to right. The curve begins in its lower-left corner with screener values that represent no true-positives and no false-positives (i.e.,  $TPF = FPF = 0$ ) and proceeds to the upper-right corner where all cases are true-positives or false-negatives (i.e.,  $TPF = FPF = 1$ ).

ROC curves are convex and appear in the upper-left half of the unit-square with a decreasing slope. Figure 8.2 shows two hypothetical curves. A curve that represents a screener that carries no information (i.e., decisions at chance) would lie along the diagonal from the lower left to upper right ( $TPF \approx FPF$ ). For a useless screener like this one, the two populations shown in Fig. 8.1 would overlap perfectly. For an accurate screener, the curve will start in the lower-left corner where TPF and FPF both equal .00; the curve will increase rapidly to represent relatively poor diagnostic utility.

Such idly while FPF values are still low and then level off after TPF nears 1.00; the curve will end in the upper-right corner with TPF and FPF at 1.00. Such a perfect screener would produce distributions in Fig. 8.1 that are separated entirely by at least one score on the screener. The two curves in Fig. 8.2 (i.e.,  $A \approx .70$  and  $A \approx .90$ ) fall between the two extremes of perfect prediction and uselessness.

The TPF and FPF can be described in terms of conditional probabilities, which require the definition of some terms. Membership in the difficulty population, which in education is defined by a criterion test's standard for proficiency, is labeled  $D=1$ , and membership in the typically achieving population is labeled  $D=0$ . We use the term "difficulty" primarily as a mnemonic for the meaning of  $D$ . A screener can then have a positive test result,  $Y=1$ , which indicates the likely membership in the difficulty population (i.e., left of the vertical line in Fig. 8.1), or a negative test result,  $Y=0$ , which indicates the likely membership in the typically achieving population. We define the TPF as the probability of a positive screener result—an indication of a likely reading difficulty—among students who fall into the reading difficulty population:  $P(Y=1|D=1)$ . The TPF is frequently called *sensitivity*, which refers to how acutely a screener can detect students with true reading difficulties. Some authors refer to TPF as the true-positive proportion or rate.

The FPF equals the probability of a positive screener result among students who achieve an acceptable standard on the criterion measure:  $P(Y=1|D=0)$ . The FPF quantifies the likelihood that the positive test was actually false. The FPF has also been labeled the false-positive proportion or rate. Rather than plotting TPF against FPF, many sources refer to replace FPF with  $1 - \textit{specificity}$ . Specificity equals  $1 - \text{FPF}$  or  $P(Y=0|D=0)$ , the true-negative fraction. It describes how well a screener can rule out unwanted cases. A *specific* screener minimizes the number of students who test positive on the screener in error. The ROC curve is often shown as a plot of sensitivity on the vertical axis versus  $1 - \textit{specificity}$  on the horizontal axis, but this is the same as a plot of the TPF versus the FPF. We will use both sets of terms, as they are common in the literature and each has benefits in different contexts.

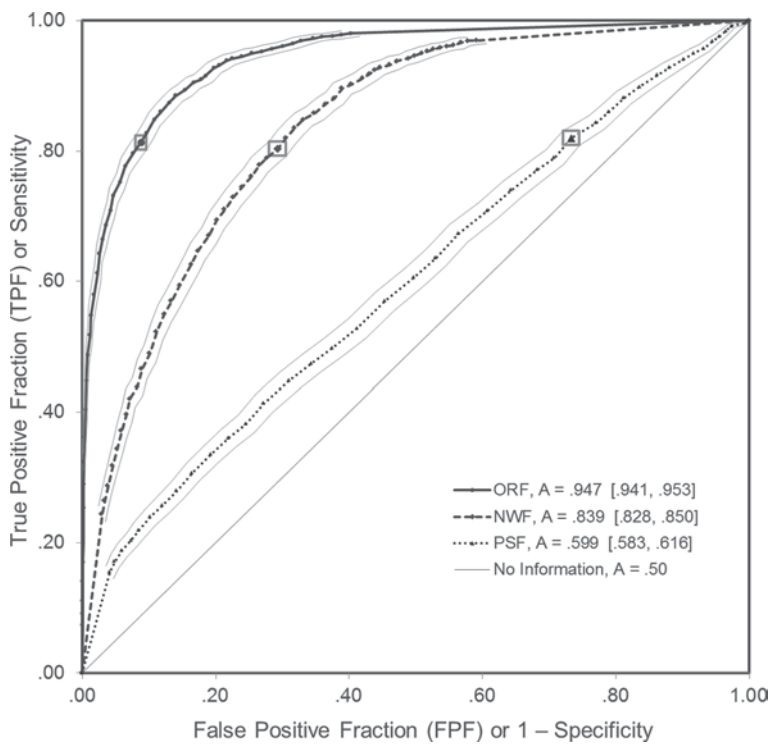
The area under the ROC curve,  $A$  (as illustrated in Fig. 8.2), usefully summarizes the overall performance of a screener across a range of possible decision thresholds and represents the mean sensitivity or TPF averaged uniformly over the range of specificity or FPF values and vice versa;  $A$  describes the average sensitivity over all values of specificity.  $A$  also has a useful interpretation in a forced-choice task, where a decision maker is provided with scores from two individuals, one selected at random from the reading difficulty population and one selected randomly from the population of typically achieving students. In this scenario,  $A$  represents the likelihood that the two students will be correctly assigned to the two populations based on their scores on the screener. Importantly, as a measure of accuracy,  $A$  is not confounded by either the proportion of students in the reading difficulty population, termed the *base rate*, or the specific decision threshold (cut score) chosen for the screener.

As noted above, values of  $A$  may range from .50 to 1.00. The ROC curve for a screener with no information follows the diagonal, so the value of  $A$  for such a measure would be about .50. With a useless screener, the ability to classify students

in the forced-choice task would be no better than the flip of a coin. A nearly perfect screener would lead to a curve that covers almost the entire area of the unit square. In this case,  $A \approx 1.00$  and implies nearly perfect assignment of any a pass–fail pair of students in the forced-choice task. Precise criterion values for the area under the ROC curve have not been established, but generally speaking, “values of  $A$  between .50 and .70 or so represent a rather low accuracy—the true-positive proportion is not much greater than the false-positive proportion anywhere along the curve. Values of  $A$  between about .70 and .90 represent accuracies that are useful for some purposes, and higher values represent a rather high accuracy” (Swets, 1988, p. 1292). For many purposes, values of  $A$  above .95 indicate an excellent screener, .85–.95 signify a very good screener, values .75–.85 suggest a reasonable screener. Values below .75 represent relatively poor diagnostic utility. Such low accuracies may have their greatest value in cases where judgments with other methods are very poor and the consequences of the incorrect choice can be especially costly or dangerous (e.g., invasive surgery). In reading instruction, teachers have a reasonable capacity to judge student performance (Martin & Shapiro, 2011), and for values of  $A$  below .75, we believe their judgments are likely more valuable than a weak screener.

Figure 8.3 shows three ROC curves, one each for three DIBELS 6th Edition screening measures administered in the spring of first grade, ORF, nonsense word fluency (NWF), and phoneme segmentation fluency (PSF), with a sample of students fluent in English (Smolkowski & Cummings, 2014). Each of these measures was used to discriminate between students in a population that is at risk for reading difficulty, defined by those below the 20th percentile on the *Stanford Achievement Test—10th Edition* (SAT10; Pearson Education, Inc., 2007), from those not at risk for later reading difficulty with a sample of 4885 students.

For each measure, the figure shows the plot of the TPF against the FPF along with two thin, gray lines that show the confidence bounds around the TPF and FPF at each point. From the description of the ROC curves and  $A$ , it is clear that the curve for ORF, with  $A = .95$ , represents a fairly accurate screener. Given ORF scores from two students, one from the reading difficulty population and one from the typically achieving population, the ORF at the end of Grade 1 would correctly order those two students 96% of the time. The curve for NWF,  $A = .84$ , is less accurate, but NWF may still be useful for instructional decisions, especially as it focuses on a specific subskill (decoding). The curve for PSF, however, might be characterized as mediocre or poor, as  $A = .60$  implies that a decisions based on that measure at the end of first grade would improve only marginally on chance. At that level of accuracy, teachers are likely to be better judges of student performance and its relevance to student membership into the two assumed underlying populations. Due to the large sample size ( $N \approx 4885$ ) in this example, statistical uncertainty was very low. The 95% confidence bounds for  $A$  were .94–.95 for ORF, .83–.85 for NWF, and .58–.62 for PSF. For each measure, Fig. 8.3 also shows a larger point on the ROC curve surrounded by a small box, which represents the decision threshold for the lowest score that exceeds a TPF value of .80. The small box shows the confidence bounds on the TPF and FPF at that cut point.



**Fig. 8.3** ROC curves for oral reading fluency (ORF), nonsense word fluency (NWF), and phoneme segmentation fluency (PSF) from the spring administration of first grade used to discriminate between students at risk and those not at risk determined by the 20th percentile on the SAT10. The area under the curve,  $A$ , indicates that ORF was quite accurate, NWF was less accurate but still useful, and PSF was poor. Values of  $A$  in brackets represent 95% confidence bounds, and the light lines surrounding the curve for each measure show 95% confidence interval for TPF and FPF at each potential cut point. The large markers on each line, surrounded by a small box, indicate the decision thresholds where the TPF (sensitivity) exceeds .80: 31 correct words per minute for ORF, 62 correct letter sounds for NWF, and 61 correct phonemes for PSF. The small box shows the 95% confidence bounds on TPF and FPF at the selected decision thresholds

Several other indices to assess accuracy have been suggested, but most have been dispensed as inadequate to the task due to their dependence on either the base rate or decision mechanism (Swets, 1986), which can lead to misleading results. The overall percentage correctly classified by the screener, for example, depends on both the decision threshold and the base rate of the outcome of interest in the chosen sample and can be biased. Sensitivity, by itself, is also inadequate (Swets, 1988). Sensitivity depends on the chosen decision threshold and associated level of specificity and cannot solely describe the accuracy of a screener. Specificity similarly relies on a level of sensitivity given a chosen cut score. Even when provided as a pair, sensitivity and specificity offer a limited description of accuracy because they depend on the chosen decision threshold. This makes comparisons between screeners particularly difficult as different thresholds for any given screener will produce different pairs of sensitivity and specificity values. Comparisons of screen-

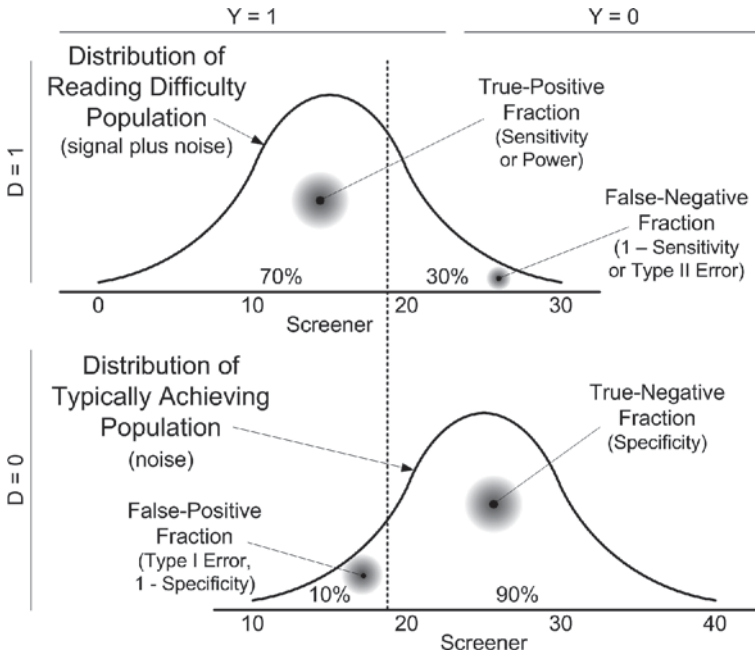
ers based upon threshold-specific statistics is similar to comparing the capacity of two containers that contain some prespecified amount of water rather than when filled to their maximum volume. In contrast, the ROC curves in Fig. 8.3 allow the direct comparison of measures across their entire range of scores. Similarly,  $A$  summarizes the curves in a way that also allows comparisons across measures without relying on a single decision threshold.

Some authors have proposed that predictive values can be used to assess accuracy, but these statistics are also problematic. The positive predictive value (PPV) characterizes the probability of failure on the criterion test among those who test positive on screener,  $P(D=1 | Y=1)$ , and the negative predictive power (NPV) describes the probability of passing the criterion test among those who test negative on screener,  $P(D=0 | Y=0)$ . While these indices have their uses, like TPF and FPF, they also depend on the chosen decision threshold (e.g., benchmark goal or cut score), which makes them a poor measure of the *overall accuracy* of a decision system. Moreover, predictive values depend on the base rate on the outcome event in the sample in which they were derived, which violates one of the fundamental attributes of a measure of accuracy (Swets, 1988). With a perfect screener, for example,  $PPV=NPV=1.0$ , but for a screener that carries no information, PPV will equal the base rate,  $\rho$ , and NPV will equal its complement,  $1-\rho$ . Unlike sensitivity and specificity, the predictive values seldom cover the range from 0 to 1, which means they cannot be used to create a curve similar to the ROC, and the area under such a curve would not be defined for the entire unit square.

The accuracy of the diagnostic system provided by the ROC curve and  $A$  offer valuable information, such as whether one screener is inherently better at classifying students in the two underlying populations across the range of values. The ROC and  $A$  also provide information about whether, for a particular scenario, a different cut score might offer a better trade-off between sensitivity and specificity (or between predictive values). Many approaches to accuracy other than the ROC curve and its summary statistic,  $A$ , fail to address the full range of screener values and its overall worth as an accurate predictor of performance, such as the relying on a statistics from a single contingency table (e.g., sensitivity, NPV, percent correctly classified). “In short, indices defined in terms of a single  $2 \times 2$  table confound discrimination capacity and decision criterion” (Swets, 1986), and reporting only such measures is analogous to describing a randomized trial only for subjects with a single pretest score. The selection of the area under the ROC curve,  $A$ , as the ideal index of accuracy (Swets, 1996), however, did not stem from the problems with other proposed measures of accuracy. Rather, these methods have been formulated from a rich foundation in statistical theory.

## ***Signal Detection Theory***

Signal detection theory and the methods for the evaluation of diagnostic systems were originally developed from work on hypothesis testing (e.g., Neyman & Person, 1933). “Modern detection theory treats the problem as one of distinguish-



**Fig. 8.4** Overlapping populations for typically achieving students (noise) and students with reading difficulties (signal + noise). The figure depicts the four outcomes with associated terms used in signal detection and null-hypothesis testing theories. In signal detection theory, the ROC curve represents the relative values of the TPF and FPF as the vertical line moves from left to right

ing between two statistical hypotheses” (Swets, 1988, p. 1291; see also Peterson, Birdsall, & Fox, 1954; Streiner & Cairney, 2007). The challenge involved the discrimination between a noise distribution and a signal-plus-noise distribution. Early researchers noticed that their problem was not unlike those addressed by the null-hypothesis testing framework, where investigators compare a distribution under the null hypothesis (noise) to one that includes an intervention (signal plus noise). They began to consider two overlapping normal curves and the choice of a critical value to achieve certain levels of sensitivity (TPF) and specificity (1–FPF) values.

The four outcomes in signal detection theory are identical to those used in hypothesis testing (see Fig. 8.4). Specificity equals the true-negative fraction. Its complement (1–specificity) equals the FPF, which is called the Type I error rate in hypothesis testing. The false-negative fraction is equivalent to the Type II error rate in hypothesis testing. Its complement, the TPF (sensitivity), is equivalent to statistical power in the null-hypothesis testing framework. Hence, the ROC curve plots the relative values of the two operating characteristics for null-hypothesis testing, power and the Type I error rate, on the unit square. Figure 8.4 depicts a decision threshold with the TPF or power at 70% and the FPF or Type I error rate at 10%. One can also draw direct comparisons between the area under the ROC curve, *A*, values of Cohen’s *d*, the point-biserial correlation coefficient, and other measures of effect size (Swets, 1996; Rice & Harris, 2005). Like *A*, Cohen’s *d*, for example, also

characterizes the separation between two populations. Chapter 1 in Swets (1996; also Swets, 1973) provides an extensive history of the development of ROC curves and methods for the evaluation of diagnostic systems.

### The Assumption of Two Populations

As detailed above, signal detection theory relies fundamentally on the assumption that individuals are placed into two populations. In education, as in many fields, the two populations are typically created by using a comprehensive criterion measure, such as a state reading test or the SAT10. The criterion places students into populations: one for students with acceptable performance on the test and one for unacceptable performers. This assignment by a criterion test can be somewhat arbitrary, but it is necessary to meet the assumptions of the methods. Although this criterion is often called a *gold standard*, the implied perfection is generally unattainable. Few fields have even a near-perfect, reproducible criterion (Swets, 1988). When testing a cancer screener (e.g., mammogram), biopsies may be used as a criterion measure, but surgeons may miss the affected tissue. In engineering, destructive stress tests may be used as a criterion (e.g., tests of airplane wing failure), but such tests can seldom duplicate all the forces at work during flight. Imperfect criterion tests are actually quite common across fields, and the reproducibility of results is “not uniformly higher for the diagnoses based on ‘hard’ rather than ‘soft’ evidence” (Kraemer, 1992, p. 15). The lack of a perfect measure will most likely depress estimates of accuracy, so the area under an ROC curve may never feasibly reach 1.0 for criterion measures, such as those in education and psychology that typically have some flaws. Kraemer (1992) suggests methods for characterizing the reproducibility of criterion tests, such as with kappa, and portraying their magnitude (e.g., Landis & Koch, 1977). Nonetheless, diagnostic systems frequently outperform alternative approaches to decision-making even with imperfect criterion measures (Swets, 1996; Swets et al., 2000a, 2000b).

An adequate criterion measure should meet several conditions. It should be a reliable, valid, and accurate indicator of the content under investigation. If the criterion measure is ordinal or continuous, it should have a value that justifiably places test cases into the two populations in order to gauge the value of a screener. Most importantly, the determination of the population memberships should not depend on either the system under evaluation or the test sample. “The truth about sample items should be determined without regard to the system’s operation, that is, without regard to the system’s decisions about test cases” (Swets, 1988, p. 1290). Evaluating a screener or screening system against itself will inflate the apparent accuracy (Swets et al., 2000b). Similarly, the procedures to establish whether an event has occurred should not affect the sample under investigation, such as the evaluation of a classification system that determines acceptance into college that relies on a sample of students accepted into college. Finally, a test sample, used for the evaluation of a screener, should include cases that represent the different types of decisions and adverse events, the four outcomes, to which the system should be applied.

Many measures are adequate for the evaluation of literacy screeners. Criterion measures differ but frequently involve validated, norm-referenced tests that allow



placement of students into populations based on a percentile. Petscher, Kim, and Foorman (2011) used the SAT10 and *Gates-MacGinitie Reading Test—Fourth Edition* (MacGinitie & MacGinitie, 2006) as criterion measures. Silbergitt and Hintze (2005) reported on ORF from the winter of 1st grade through the spring of 3rd grade, and all ORF assessments were evaluated against the reading portion of the *Minnesota Comprehensive Assessment*. Hintze, Ryan, and Stoner (2003) used the CTOPP as their criterion, which they collected at the same time as the DIBELS screener measures they evaluated. Smolkowski and Cummings (2014) and Smolkowski, Cummings, and Baker (2014) used the SAT10 as the criterion measure for Grades K through 2 and the OAKS for Grade 3. Smolkowski and colleagues chose the 20th normative percentile on the SAT10 as the criterion for high risk and the 40th normative percentile to distinguish some risk from students at acceptable levels of performance (benchmark). The choice of criterion was selected more so by convention than an established standard. They characterize students below the 20th percentile as at *high risk* and those below the 40th percentile as below benchmark (or some risk).

When choosing a criterion measure or interpreting the results of a published study, it is important to consider what skill or ability is assessed by the criterion measure. An investigator should choose a criterion measure that defines the population of students at risk in similar terms as the screener under investigation. Similarly, educators should choose a screener that has been evaluated with respect their needs, or at least a reasonable approximation thereof. For example, the analysis of Smolkowski and Cummings (2014) may work well in a school that uses the SAT10 or a state test that correlates highly with the SAT10 but possibly not in a school that uses a very different standard. Thus, like any research study, the details, such as the criterion test, the chosen criterion level to describe risk, and the sample of students selected all help determine the generalizability of the study and its applicability within any other given context.

**Decision Thresholds** Once a screener has been shown to be accurate with respect to an appropriate criterion measure, its practical use depends on the choice of a *decision threshold*, a score on the screener that best discriminates members of the two populations of interest. For literacy instruction, the decision threshold specifies the score on the screener below which students are more likely to come from the reading difficulty population and less likely to come from the typically achieving population. In practice the decision threshold is the point on the screener that determines whether the student receives additional support. It is important to recognize that the decision threshold is not a mandate to provide services. As with any assessment, measurement error is always present. Thus, for some students who fall just below the decision threshold, their teachers may choose to retest them, test them with a more sensitive measure, or more closely observe their progress in Tier 1 before making any changes to their instructional program. We recommend data-based decisions, however, whenever a good screener or other assessment is available, because decisions based on clinical or professional judgment rarely outperform decisions based on statistically generated rules (Dana & Dawes, 2004; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954).

Several procedures can be used to produce a valid decision threshold, which reiterates the importance of prioritizing at the outset of a study the decision rules that will be used to select a given cut score. The particular procedure should be in line with the theoretical construct of the difficulty wished to diagnose and the overall goals of the diagnostic system (Pepe, 2003). Most procedures depend on the use of a  $2 \times 2$  contingency table, where one dimension represents the criterion (i.e., reading difficulty vs. typically achieving students) and the other a particular score on the screener. Each score on the screener—each potential decision threshold—produces a different table, and the table can be summarized with a wide array of statistics (Streiner, 2003; Schatschneider, Petscher, & Williams, 2008). From these statistics, we discuss a subset of the possible methods to select a decision threshold.

Swets et al. (2000b) discuss a general decision method that attempts to “maximize the ‘expected value’ of a decision, i.e., to maximize its payoff in the currency of benefits and costs” (p. 9). They express the decision goal as a formula in terms of the probabilities of true population membership, such as the typically achieving or reading difficulty populations, and the benefits and costs associated with the joint occurrence of a particular population membership and screener decision. When the screener decision and population memberships agree (i.e., true-positives and true-negatives), the choice results in benefits. When they disagree (i.e., false-positives and false-negatives), the choice results in costs. The formula and procedure articulated by Swets et al. (2000b) works well when costs and benefits can be estimated clearly and accurately, which offer a clear justification for the decision rule.

In most educational situations, useful approximations of costs and benefits have been difficult to obtain. In such cases, the selection of a decision threshold based on sensitivity and specificity offers a reasonable alternative. Swets et al. (2000a), for example, suggest that a simple goal might maximize the sensitivity for a given level of sensitivity or specificity. The content area and context intended for a diagnostic system will help determine the decision rule used to set a decision threshold. As with most analytical approaches, it is important to select a decision rule for the selection of a cut score a priori to avoid post hoc choices that may be sample specific or biased by other factors. We review a sample of approaches to the selection of a decision threshold below and offer an example justification for one potential rule, based on sensitivity.

## Sensitivity and Specificity

Sensitivity and specificity offer one method for the selection of a decision threshold. The cut score may be chosen to meet a defined level of specificity, a particular level of sensitivity, or some combination. We next discuss one example of a decision rule based on sensitivity within a prevention-oriented screening model, which could be adjusted or changed to fit other contexts in education.

The selection of a decision threshold for reading fluency measures within a prevention-oriented approach suggests a focus on sensitivity, or more precisely, the complement of sensitivity, the false-negative fraction. A rule based on sensitivity prioritizes the identification of students who will perform below a given

standard determined by the criterion measure. A careful consideration of the goals surrounding reading instruction might lead to the rule that a reading fluency screener should incorrectly identify *not more than 20%* of students from the reading difficulty population as typically achieving. Relying on this premise, Smolkowski and Cummings (2014) and Smolkowski et al. (2014) decided to select decision thresholds corresponding to the score on each DIBELS measure with a false-negative fraction at or below .20. This translates into the score where sensitivity is equal or greater than .80, which would correctly classify 80% of struggling readers.

This decision rule for the selection of a cut score has several benefits. First, false-negatives appear to be a more problematic error when it comes to educational decision-making than false-positives. False-negatives represent students who will likely not receive additional instructional supports or interventions even though they may need such supports. They have been deemed a typically achieving student who happens to fall in the lower end of the distribution. In contrast, false-positive students will likely receive an intervention, additional supports, or at a minimum, additional monitoring. Providing additional instruction to students who do not necessarily need it seems more ethically responsible than the failure to provide such instruction to a student who truly needs it. Second, false-negative errors are more difficult to correct than false-positive errors. Because teachers are less likely to monitor or intervene with students who screen negative, teachers are also less likely to notice prediction errors when they occur. A false-negative student will not likely receive the additional attention from his or her teacher or instructional assistant because she will not receive additional services. Such a student may flounder in whole-class instruction for some time. In contrast, teachers are more likely to identify those students who screen positive in error, so teachers can take corrective action more quickly. If a typically achieving student performed poorly on a screening assessment (e.g., ORF) and becomes incorrectly assigned to small-group instruction, his teacher or instructional assistant may quickly determine that he does not need to participate in small group instruction. That student could be placed back in standard instruction at any time (e.g., in Tier 1 in a response-to-intervention framework).

As the purpose of DIBELS, as a diagnostic system, is to identify students with the potential for reading difficulty or skill deficits, we chose to focus our criteria on the statistics associated with the population of students potentially with reading difficulty rather than the population of typically achieving students. That justification applies well to reading fluency measures in the context of a multitiered system of instructional supports, but it does not represent the only decision rule. Within similar contexts, a more stringent decision rule might be chosen to correctly classify 90% of struggling readers (sensitivity  $\geq .90$ ). A similar approach and justification for a different diagnostic system might rely on specificity or its complement (the false-positive fractions) if such decision rules met the goals and context for the system.

## **Youden Index**

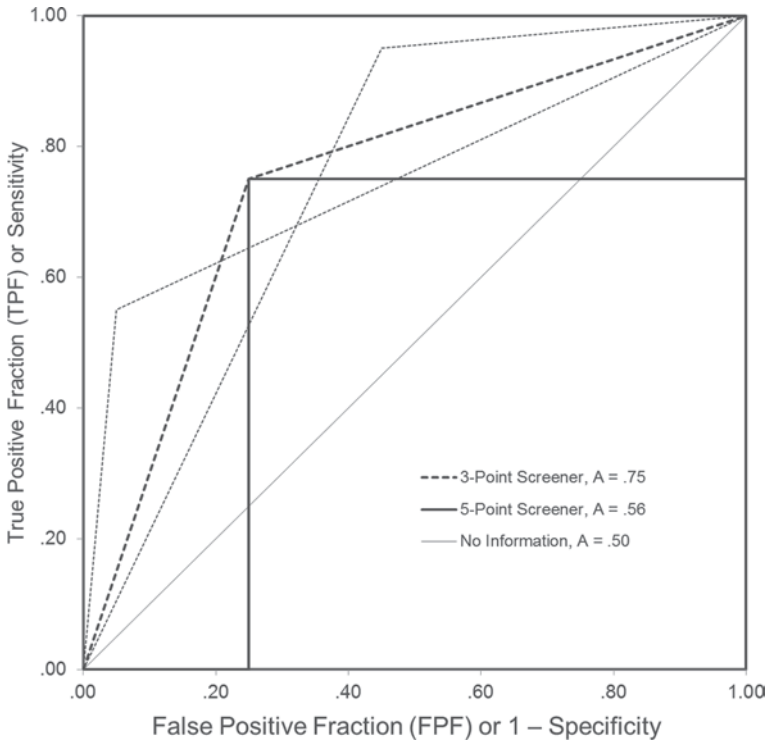
In some situations, or perhaps if a choice between reliance on sensitivity or specificity is unclear, a rule based on a combination of the two might be most appropriate.

Youden (1950) described a procedure for selecting a decision threshold that maximizes the combination of sensitivity and specificity. To calculate the Youden index, one simply computes  $J = \text{sensitivity} + \text{specificity} - 1$  (equivalently,  $J = \text{TPF} - \text{FPF}$ ) for each score on the screener. The decision threshold is the score that maximizes  $J$ . This process typically selects the point on the ROC curve closest to the upper-left corner and produces results similar to those from the more complicated, iterative procedure recommended by Silbergitt and Hintze (2005, see p. 316). The Youden index implies that the costs of the errors associated with sensitivity and specificity are weighted equally as it simultaneously maximizes them. In contexts where the two types of errors have similarly weighted consequences, the largest value from the Youden index may usefully determine the decision threshold. Conversely, when the two errors carry very different costs, such as the identification of students with the potential for reading failure, this equal weighting of true-positives and false-positives may make it a poor choice for the selection of a decision threshold.

### Boundary Scenarios

Nearly any screener can, at some point, attain a sensitivity value greater than say, .80. But at the same decision threshold, less accurate screeners may produce very low specificity. In the framework presented herein, this concern is minimized because it is unlikely that a measure with such discrepant sensitivity and specificity values would have achieved adequate accuracy (i.e.,  $A$  is likely poor). Decision thresholds are typically only useful to generate given reasonably accurate screeners or diagnostic tests. Figure 8.3 offers an example of a measure, PSF, that demonstrates the scenario of a screener with low accuracy. At the score (61) where PSF attains a sensitivity value greater than .80 (.82), specificity equals .27. This indicates a very high rate of false-positives, 73%. A decision threshold that correctly classifies only 27% of the students in the typically achieving population is likely unacceptable for many instructional situations. We could have anticipated a result like this one because PSF had a very low accuracy value,  $A = .60$ . Screeners with very low accuracy often result in decision thresholds with discrepant sensitivity and specificity values, and we would not recommend selecting a decision threshold for measures such as these. The use of signal detection theory is the first to demonstrate sufficient accuracy of a measure generally rules out choosing decision thresholds for inaccurate measures that could lead to an unsatisfactorily low or discrepant sensitivity and specificity values.

A potentially interesting theoretical argument is whether measures could be designed to have excellent discrimination around a single point yet no other points, suggesting that a measure could have high sensitivity and specificity values (e.g.,  $\geq .75$ ) yet not achieve a reasonable accuracy level (e.g.,  $A < .75$ ). We illustrate two example cases that suggest this scenario is highly unlikely. In the first scenario, assume a 3-point screener that produces sensitivity and specificity values equal to .75 at the midpoint, with the boundary values at each extreme (e.g., sensitivity=0 and specificity=1). Such a screener will have an accuracy value of  $A = .75$ . Similarly, other 3-point screeners with  $A = .75$  might have sensitivity–specificity pairs of .55 and .95 or .95 and .55 at their midpoints, respectively. In Fig. 8.5 we display



**Fig. 8.5** ROC curves for three 3-point screeners (*dashed lines*) and one 5-point screener (*solid line*). All three 3-point screeners produce accuracy values,  $A$ , of .75. The *heavy dashed line* represents a screener with sensitivity and specificity both equal to .75. The *light dashed lines* represent screeners with sensitivity and specificity equal to .55 and .95 and vice versa. The *solid line* represents a 5-point screener with maximal discriminability at its upper-left most point, where sensitivity and specificity both are equal to .75. This screener, however, has an accuracy value,  $A = .56$ , barely above the no-information screener, due to the other points in the curve, where either sensitivity or specificity is equal to zero

the ROC curves associated with these three hypothetical 3-point screener scenarios (*dashed lines*). Each of the three 3-point screeners achieves an acceptable level of accuracy,  $A$ . If additional scores were added to one of the screeners and the curve maintained the concave nature typical for ROC curves, it would then produce accuracy values at or above .75. These hypothetical scenarios suggest that it is unlikely that a measure could have high sensitivity and specificity values around a single point yet not achieve a reasonable accuracy value,  $A$ .

As a second example, we present a 5-point screener that has excellent screener properties at a single point, yet no discriminative properties at four points. That is, at the decision threshold the screener achieves sensitivity and specificity values equal to .75, yet at all other points, either sensitivity equals zero with specificity equal to .75 or specificity equals zero with sensitivity equal to .75. This type of screener would not maintain a concave ROC curve. Rather, it would have convex sections. Figure 8.5 also shows this scenario with the dark solid line (square).

This curve produces a much lower accuracy value,  $A = .56$ . In the forced-choice task mentioned earlier, this indicates that the screener, as a whole, would only correctly classify students from the two populations about half the time (56%), barely better than chance. Given the adequate sensitivity and specificity values, how can this be?

The answer has to do with the other cut points on the measure. Cut points above or below the decision threshold provide no discrimination in one characteristic dimension or the other. Such a 5-point reading screener has sensitivity and specificity values of .00 and 1.00 at a score of 0, .00 and .75 at a score of 1, .75 and .75 at a score of 2, and .75 and .00 at a score of 3, and 1.00 and .00 at a score of 4. As a consequence, a score of 1 has a true-positive fraction of zero, or no discrimination among the reading difficulty population. Similarly, a score of 3 does not discriminate among the typically achieving population. Such a screener, however, would produce very unlikely, bimodal distribution of the two populations, which becomes clear when considering one population at a time. Considering false-positives ( $1 - \text{sensitivity}$ ) of the typically achieving population, no members scored 0, 25% scored 1, none scored 2, 75% scored 3, and none scored 4. Turning to true-positives (sensitivity) in the reading difficulty population, none scored 0 or 1, 75% scored 2, none scored 3, and 25% scored 4.

The scenarios above considered two general types of screeners optimized for single cut scores. We believe the examples show how  $A$  is valuable as a measure of overall accuracy even for single-score-optimized screeners. In contrast, because  $A$  is a function of sensitivity and specificity, it is not likely to find a screener with acceptable accuracy yet no cut score with potentially adequate sensitivity and specificity values.

*Predictive Values* The use of predictive values to establish decision thresholds has also been put forth, primarily in the education literature base, as a recommended practice. This recommendation leads to a number of challenges because predictive values are not indices of the accuracy of diagnostic tests (Pepe, 2003; Swets, 1996). Rather, uncorrected predictive values depend on the base rate of reading difficulty in a given sample and thus serve to characterize the clinical or educational significance of the test in a manner that is analogous to a local normative comparison (versus a national norm comparison). Although predictive values can be useful for teachers or parents, because they are sensitive to the base rate on the criterion test, they do not characterize the accuracy of the test and are not useful for selecting decision thresholds on screener measures that will persist across other samples of schools. "A low PPV may simply be a result of low prevalence of [reading difficulties] or it may be due to a [screener] that does not reflect the true [reading difficulty] status of the subject very well" (Pepe, 2003, p. 16).

It might be helpful to define PPV in terms of Fig. 8.4. PPV is the relative proportion of students to the left of the dashed vertical line who are in the reading difficulty population rather than the typically achieving population. Within signal detection theory, the PPV does not portray the discrimination between two populations because it represents a relative number of students within just a portion of each of



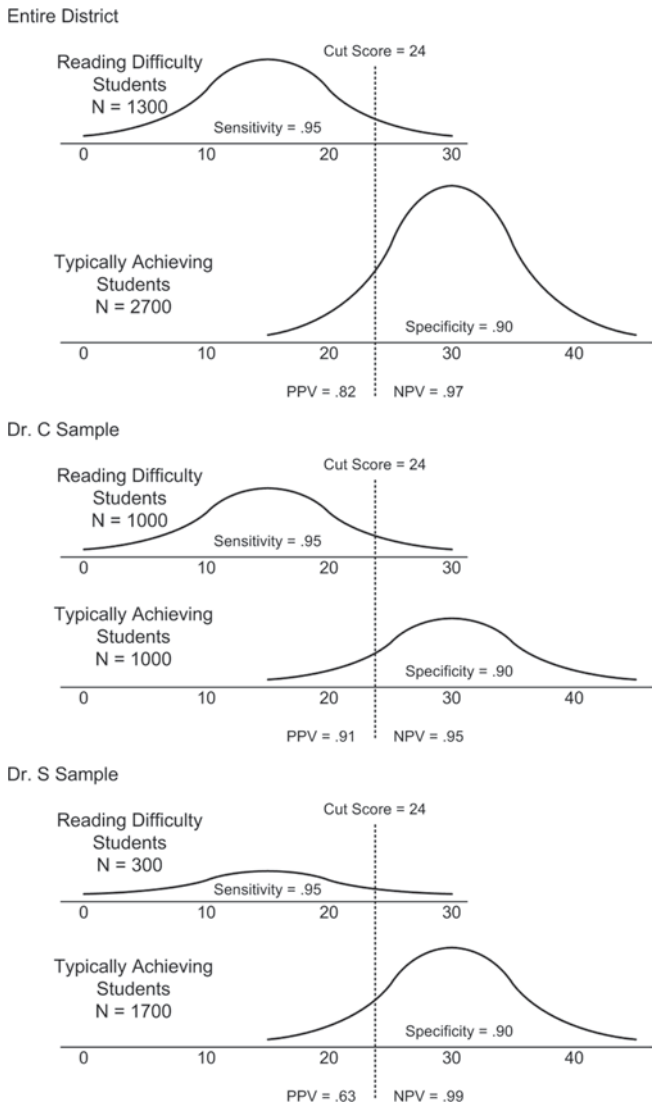
the two populations. The proportion of students to the left of the decision threshold depends on the number of students in the each of the two populations, the shape of the two distributions, and the choice of the decision threshold. As described above, in a screener that carries no information, the PPV will equal the base rate,  $\rho$ , so the PPV for any given screener can range only between  $\rho$  and 1.0.

To further demonstrate the challenges with predictive values, consider a thought experiment. Dr. C and Dr. S plan to evaluate the same reading screener against the same criterion measure, and each researcher selects a sample of students from the same district. Both researchers want to know how well the chosen cut score of 24 appropriately selects students who are at risk. The district as a whole has 1300 total students with reading difficulties and 2700 students who perform adequately. Dr. C and Dr. S, however, have different sampling goals in mind, so Dr. C sampled the same number of students from each of the reading difficulty and typically achieving populations (1000 students each). Dr. S selected approximately 23% of the reading difficulty population (300 students) and about 63% of the typical achievers (1700 students). To avoid the potential for bias, each of the researchers randomly sampled the populations of reading difficulty and typically achieving students, respectively. That is, the means and standard deviations for the samples collected by Dr. C and Dr. S match those for the whole district. The distribution of the reading difficulty populations and typically achieving populations for the whole district and each of the two samples are depicted in Fig. 8.6.

Because the two samples of students have the same distributions as the entire student enrollment in the district, Dr. C and Dr. S estimate nearly identical sensitivity and specificity values, about .95 and .90, respectively. Both samples would also produce the same ROC curve and  $A$  values as an analysis of all students in the district, except for some random sampling error. The proportion of students in each sample differs—Dr. C included nearly 77% of all reading difficulty students, whereas Dr. S included just 23%—so their predictive values will also differ. Dr. C calculated  $PPV = .91$  and  $NPV = .95$  but Dr. S obtained  $PPV = .63$  and  $NPV = .99$ . These values, when calculated for the whole district, are .82 and .97, respectively. As the base rates differed in the two samples, so will the investigators' conclusions about the value of the screener, especially if they rely on PPV. As described earlier, both predictive values depend on the base rate, and for Dr. C the base rate is 50%, so her predictive values can range from .50 to 1. For Dr. S, however, the base rate is .15, so  $PPV \geq .15$  and  $NPV \geq .85$ .

Had both investigators drawn conclusions about the accuracy of the chosen cut point from their respective samples using predictive values, they would have drawn considerably different conclusions even if they used the same a priori decision rules. Let us say for example that Dr. C and Dr. S agreed at the outset of the study that their aim was to accept the cut point as valid if  $PPV > .80$ . This decision rule is akin to noting that at least 80% of the students who screen positive (i.e., at risk for later reading outcomes) will actually perform below the criterion standard. Given that PPV will range from .15 to 1 in Dr. S's sample, he is much less likely to find the cut point of 24 acceptable compared with Dr. C.





**Fig. 8.6** Three sets of overlapping distributions from a hypothetical school district: the district population, a sample collected by Dr. C, and a sample collected by Dr. S. In each case, the means and standard deviations for reading difficulty students are assumed to be the same, similarly for the means and standard deviations of the typically achieving students

This thought experiment is clearly hypothetical, but it represents real studies in some respects. Often the statistical evaluation of screeners relies on samples that were collected for other purposes. For example, the sample in Smolkowski and Cummings (2014) relied on students in schools who participated in Oregon Reading First. Due to the Reading First participation criteria for districts and schools, the

original sample likely included more struggling readers than many schools; it may be more similar to the Dr. C sample. Nonetheless, because they relied on specificity to select decision thresholds, their evaluation is not subject to the same limits as an evaluation based on predictive values. This is not to say that the sample of Oregon Reading First Schools generalizes to all other students. For example, to generalize to a new set of schools, the shape of the distribution for students with reading difficulties in Oregon Reading First Schools should be similar to the shape of the distribution within the new set of schools. But the shape of the distribution is more stable than the base rate, since the base rate is a function of both the sample sizes selected for an evaluation as well as the shape of the underlying distributions.

Thus, all evaluations of diagnostic systems may generalize to only a subset of students and schools. To generalize decision thresholds based on predictive values, however, investigators have an additional burden because they need to demonstrate that not only does the sample have similar shaped distribution—similar means and standard deviations at a minimum—but they also need to ensure that their sample of students has the same relative proportions as the population of students with reading difficulties and the population of typically achieving students to which they hope to generalize. Unfortunately, in most cases such validating data are unavailable, especially for a new, untested screener. As a consequence, the selection of a decision threshold with predictive values is unlikely to be stable across samples and hence more difficult to generalize to new settings than decision thresholds selected with decision rules based on sensitivity or specificity.

Authors must also be aware of the role of base rate on their decision rules. Piasta et al. (2012) “focused on the negative predictive power when examining these results” (p. 950). One of their key measures, however, produced base rates of .14 for each of two different screeners, uppercase and lowercase letter naming. With a base rate of .14, NPV is bound by the range from .86 to 1.0 and therefore cannot address the question, “is this decision threshold good or poor?”; even a screener that carried no information (i.e.,  $A \approx .50$ ) would appear valuable in this context because it could never result in an NPV that was less than .86.

The decision rules based on predictive values need not always lead to based-rate specific cut scores. Kraemer (1992) describes a process where predictive values can be adjusted for the base rate to produce a value she calls the *quality* of the predictive value; she similarly adjusts sensitivity and specificity for the probability of a positive test. But with those adjustments, “it becomes apparent that the quality of the sensitivity is exactly equal to the quality of the predictive value of the negative test” (Kraemer, 1992, p. 99). Decision rules based on these metrics, however, take us beyond the scope of the present chapter, and other authors tend to rely more completely on sensitivity and specificity (e.g., Pepe, 2003) or more comprehensive decision rules (e.g., Swets et al., 2000a).

**Summary** The statistical evaluation of screeners requires several steps. The investigator should first choose the criterion measure that will define the two populations of interest and then select one or more screeners to evaluate in terms of their ability to discriminate between the two populations. Next, a priori decision rules should be

specified to establish the minimal level of accuracy (e.g.,  $A$ ) and the methods for the choice of a decision threshold that will determine a positive or negative test result. At this point, the investigator should test the accuracy of the screener with respect to the two underlying populations defined by the criterion measure, which requires an ROC curve and evaluation of whether the area under the curve,  $A$ , meets the decision rules. Finally, for a screener with sufficient accuracy, the decision threshold would be chosen based on the prespecified rules, such as the score where sensitivity exceeds .80 or the score that maximizes the combination of sensitivity and specificity (e.g., Youden index).

We make two recommendations for reporting the results of the statistical evaluation of screeners. First, we recommend that whenever possible, results of evaluations include ROC curves or, at a minimum, the area under the ROC curve,  $A$ , as a measure of overall accuracy. This allows for the comparison of different screeners even when investigators or educators may select different cut scores than those reported. For example, numerous papers in education discuss the diagnostic capability of individual cut scores for various screeners, yet their information is incomparable because they report statistics only for one chosen cut score. Similarly, investigators may choose to compare screeners across different learner populations (see below), where, case presenting the ROC curve, or at least reporting  $A$ , offers more complete information.

Second, as with most other types of evaluations, authors should include confidence bounds around estimates of  $A$ , sensitivity, and specificity. Few papers in education and school psychology literature report confidence intervals, which would shed light on the statistical uncertainty of estimates. For example, Piasta et al. (2012) report sensitivity and specificity values for their chosen benchmark for lowercase-letter screener of .70 and .61, but calculation of 95% confidence bounds indicates that these values could fall in the ranges of [.58, .82] and [.56, .67], respectively. Nelson (2008) similarly reports on sensitivity and specificity that, had they calculated confidence intervals, would have produced very wide bounds. In one such case, a sensitivity value estimated to be .62 had a 95% chance of a true value between .48 and .76. As with other statistical results, confidence bounds offer information about precision. When papers report single-point estimates, it suggests greater confidence than may be truly offered by the estimate. Reporting confidence bounds will provide precise estimates of precision.

One implication of the wide confidence intervals in evaluations of diagnostic accuracy is that such studies require large sample sizes. In some cases, results are based on cells sizes of one, such as a sensitivity value of .94 reported by Nelson (2008) that depended on a single student who screened positive. When interpreting results, we also recommend considering the size of the sample, and the individual cell sizes in particular, upon which the conclusions are based. When planning a study using these methods, researchers should also consider the required sample size (see Lasko, Bhagwat, Zou, & Ohno-Machodo, 2005; Malhotra & Indrayan, 2010; Pepe, 2003).

Finally, we recommend authors who evaluate the diagnostic accuracy and decision thresholds for fluency screeners consider the recommendations of the

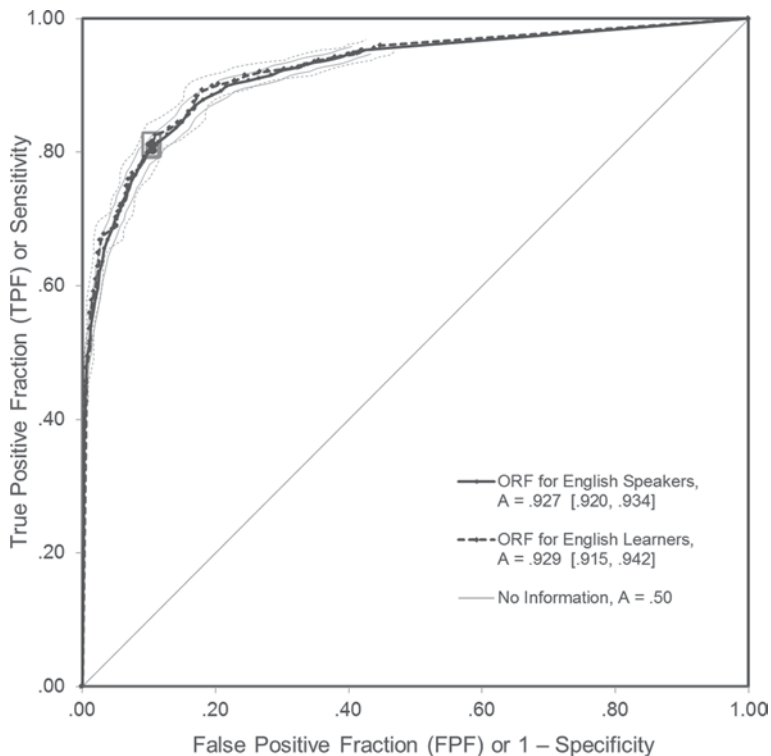
*Standards for the Reporting of Diagnostic accuracy studies* (STARD Statement, 2008; <http://www.stard-statement.org/>). The STARD Statement offers additional details that would aid in the interpretation of study results.

## Screeners in Different Populations

In this section we provide two examples that illustrate the process of selecting a cut point on a screening measure with two different general populations. Earlier, we showed how the ROC curve and associated statistics could be used to compare different fluency screeners in the same overall population of learners (e.g., Fig. 8.3). Here we show that the same methods can be used to assess the relative accuracy of a screener and its decision thresholds within different populations. We computed decision thresholds for DIBELS 6th Edition ORF at the end of Grade 1, when used to discriminate typically achieving students from those with at least *some risk* (below benchmark) for serious reading difficulty (Smolkowski & Cummings, 2014; Smolkowski et al., 2014) across two general populations namely: students proficient in the English language—most often their first language—and English learners—students who had received services for English as a second language during the year of data collection. In this example, we treated two groups of students: English learners (ELs) and English speakers (ESs) as different populations and used the same sensitivity-based decision rule to select cut scores on ORF for each group. The criterion for some risk was the 40th percentile on the SAT10, which in our sample produced base rates of .59 for ESs ( $N=4885$ ) and .83 for ELs ( $N=1960$ ).

The value of  $A=.93$  for ESs suggested that ORF discriminated students with likely reading difficulty from typically achieving students very accurately. We then calculated the optimal decision threshold using sensitivity, with the decision rule described earlier (see statistical evaluation of screeners section, sensitivity and specificity subsection). Using the sensitivity criterion, our analysis chose a decision threshold of 47 correct words per minute for ESs. The decision threshold achieved a specificity value of .89 and placed 52% of students below benchmark, indicating some risk. It was also associated with a PPV of .92, so 92% of students below the cut score were members of the reading difficulty population (i.e., performed below the 40th percentile on the SAT10). The NPV of .76 implies that 76% of ESs who screened negative were typical achievers. For ELs, ORF discriminates students with likely reading difficulty from typically achieving students with similar accuracy,  $A=.93$ , to ESs. Using the same decision rule, we selected a decision threshold of 48 correct words/min for ELs, which achieved a specificity value of .90, also similar to ESs. For ELs, however, 69% screened below benchmark, indicating some risk, and the cut score was associated with predictive values, PPV of .98 and NPV of .49.

Figure 8.7 presents ROC curves for both populations of students and shows the decision threshold on the two ROC curves. The 95% confidence bounds for  $A$  overlap, and the bounds for sensitivity and specificity at the decision threshold for ELs encompasses the decision threshold chosen for ESs. These imply that ORF, as a



**Fig. 8.7** ROC curves for ORF from the spring of first grade used to discriminate between students with some risk (benchmark) and typically achieving students determined by the 40th percentile on the SAT10. The solid curve shows the relative operating characteristics for English speakers (ESs) and the dashed curve shows the curve for students who had received services for English as a second language during the year the data was collected. Values of  $A$ , with 95% confidence bounds in brackets, indicate that ORF was quite accurate for both populations of students. The thin lines next to each curve shows 95% confidence bounds for TPF and FPF at each potential cut point. The large markers on each line, surrounded by a small box, indicate the decision thresholds where the TPF (sensitivity) exceeds .80 namely: 47 correct words/min for proficient ESs and 48 correct words per minute for English learners (ELs). The small boxes show the 95% confidence bounds on TPF and FPF at the selected decision thresholds. The boxes overlap; the bounds for ELs is larger than the bounds for ESs

fluency screener, and the associated decision thresholds perform equivalently in the two populations of students. Because the sensitivity value for a score of 47 among ELs was nearly .80 (.798), we would not likely need different thresholds for ESs and ELs, at least not for the benchmark criterion at the end of Grade 1.

A similar decision rule that relied on the NPV would have presented a different story. In the sensitivity scenario, we chose the score where sensitivity exceeded .80; so we used a rule where NPV need to exceed .80 for the selection of a decision threshold. An NPV of at least .80 selected a decision threshold of 52 correct words per minute for ESs and 82 for ELs. For ESs, the cut score of 52 was associated with

PPV of .89, sensitivity of .86, and specificity of .84, with 57% of all ESs students screened as having some risk. For ELs, the cut score of 82 was associated with PPV of .90, sensitivity of .98, and specificity of .45, and 91% of all ELs were screened as having some risk. The wide difference in the choice of decision thresholds under the NPV scenario stems largely from the greater proportion of struggling readers among ELs (83%) than among ESs (59%). The distributional characteristics are similar, however, as suggested by the similar values of  $A$ , sensitivity, and specificity at the same cut scores. When using a decision threshold based on sensitivity, the difference in base rates shows up in the proportion that screen positive, 52% of ELs versus 69% of ESs. The decision rule based on NPV, however, compounds this difference due to its dependence on the base rate and consequently screens 91% of ELs as below benchmark and 57% of ESs below benchmark. Equivalently stated, the cut point based on NPV for ELs is not specific; its specificity of .45 implies that 55% of typically achieving ELs will be incorrectly identified as having some risk.

Conceptually, characterizing a screener with predictive values is analogous to grading on a curve in that the acceptability of one student's performance depends on the performance of all other students in the classroom. This is evident in the difference in decision thresholds chosen for ESs and ELs, as their base rates differ by 24%. The choice of population decision thresholds, however, should not depend so heavily on the selected sample of students. Sensitivity and specificity, in contrast, describe the discrimination between the two underlying reading difficulty and typically achieving populations without reliance on the base rate. This is not to say that the populations do not differ in important ways that could affect their sensitivity or specificity values, but those population differences will also affect predictive values. Due to the additional challenges with base-rate dependence, we believe, however, that NPV (and PPV) are insufficient measures for the selection of decision thresholds because they lead to sample-specific decisions in potentially unpredictable ways. Decision thresholds selected to achieve a certain level of sensitivity, specificity, or some combination will tend to agree more frequently than rules based on predictive values.

## **Recommendations for the Application of Signal Detection Theory**

Signal Detection Theory and methods have a valuable place in education. With their rich history in psychology, medicine, and public health, educators have only relatively recently begun to apply the methods to educational settings. These methods can be used to set screener values, as we demonstrate in this chapter. When the methods have been appropriately applied to this purpose, researchers can also compare the value of screeners within and even across samples. Additionally, as in other fields, the application of empirical decision thresholds outperforms clinical or professional teacher judgments in terms of making special education referrals (Marston, Muyskens, Lau, & Canter, 2003). The successful use of these methods, however, depends on a thorough understanding of their development and theory.

Pepe (2003) has outlined some basic criteria for the choice of any screening or diagnostic system: (a) the difficulty, disability, or disease should have serious consequences; (b) the difficulty, disability, or disease should be relatively prevalent in the target population; (c) the difficulty, disability, or disease should be treatable; (d) the treatment should be available to anyone who screens positive; (e) the screener should not harm individuals; and (f) the screener should accurately classify individuals who have or do not have a difficulty, disability, or disease. Most educational uses fit these criteria. Nonetheless, the use of these methods could be improved.

### *Comparing Screening Systems in Education*

The evaluation framework outlined in the preceding sections recommends methods for comparing screeners and in education, perhaps more so than in other fields, researchers have developed multiple screeners for the same purpose. This naturally leads to comparisons between screeners, as educators should have the “best” screener for their intended purposes. At the outset, comparisons require first a match between the goals of the screener and the goals of an educator or researcher. But within those goals (e.g., assessment of kindergarten decoding fluency), several screeners may be available, and the methods discussed within this chapter allow for the comparison of screeners and diagnostic tests.

The ROC curve provides valuable information that researchers can use to quickly compare one screener to another. “The ROC curve transforms tests to a common scale” (Pepe, 2003, p. 72) and “by reporting the entire ROC curve, information about a test can be compared and possibly combined across studies” (Pepe, p. 71). The ROC curve provides information about the performance of the screener across the entire range of scores. A graph of the full ROC curve is particularly valuable when comparing diagnostic tests with relatively fewer values (e.g., Lewinsohn, Seeley, Roberts, & Allen, 1997; Streiner & Cairney, 2007). The area under the curve reduces the information in the ROC curve into a single number. For continuous tests with a wide range of values, such as most screeners used in classrooms, the ROC curve is relatively smooth and looks somewhat similar to the curves in Fig. 8.2 and 8.3. In our experience, *A* provides a useful summary of these screeners, although variations are sometimes required. Lasko and colleagues (2005) offer an accessible overview of the use of ROC curves, methods for graphing and computing their confidence bounds, and extensions, such as the partial area under the curve in situations when an investigator is interested in a screener only when its sensitivity exceeds a certain value.

Educators can also examine individual scores for a screener. Such evaluations do not assess the value of a screener as a whole, but only an individual decision threshold on a particular screener. Many comparisons between screeners in education, however, rely only on cut-score-specific statistics. Jenkins, Hudson, and Johnson (2007), for example, examine universal screening for reading, as screening represents the principal means to identify students who require additional intervention within a response-to-intervention (RTI) framework. The authors review studies published since 1998 and summarize the evidence on candidate measures. Jenkins and col-



leagues compare several kindergarten literacy screeners with their sensitivity values. While one screener had “poor sensitivity (50%)” (p. 588) and other had “good sensitivity (88%)” (p. 589), specificity levels “ranged from 63 to 87%” (p. 589). Each sensitivity value, however, is paired with a single specificity value and depends on the decision threshold. As demonstrated with an ROC curve (e.g., Fig. 8.3), every screener with more than a few possible scores has some point at which sensitivity equals approximately .80. A comparison similar to that made by Jenkins et al. but based on the cut scores in Fig. 8.3 would lead to the conclusion that ORF, NWF, and PSF were similarly accurate—they all have sensitivity of about .80—but their specificity ranged from .27 to .91. Figure 8.3, however, clearly shows that PSF is inferior to NWF, which is less accurate than ORF. Conclusions based on a single statistics offer little guidance, and even inferences that reference the criterion test, the chosen decision threshold, sensitivity, specificity, and other statistics offer insufficient information to make critical judgments about a screener as a whole.

In a warning about poor screeners, Glover, Albers, and Kratochwill (2007) find that “many authors have suggested that the utility of screening instruments with sensitivity, specificity, and PPVs that are below 75% (Gredler, 2000b; Kingslake, 1983) or 80% (e.g., Carran & Scott, 1992; Carter, Briggs-Gowan, & Davis, 2004; Meisels, 1989) be questioned carefully” (p. 125). A value of specificity below .75 might be well worth the trade-off to obtain a sensitivity value above .99 in the right context. Such situations arise in medicine with diseases that have a high mortality rate (e.g., HIV in the 1980s), where it is critical to identify as many individuals with the disease as possible and false-positives can be ruled out with more invasive testing. In education, there may be scenarios where educators find it more beneficial to catch all students with potential academic problems and provide intensive supports than waiting to rule out additional typically achieving students from the intervention.

Reporting and comparing the ROC curve or  $A$  resolves the challenges that stem from the reliance on sensitivity, specificity, or other diagnostic statistics (e.g., predictive values) that apply to a single value. Generally speaking a screener with a somewhat larger area under the ROC curve,  $A$ , will have, at some point in the continuum, a more useful cut score than a screener with a lower  $A$  value. Due to the relation between (a) the ROC curve and  $A$  and (b) sensitivity and specificity, the former offers a more complete and hence more valuable picture.

### **District or School-Specific Decision Thresholds**

Some authors (e.g., Richmond, 2012; Schatschneider et al., 2008) have suggested that schools use their own decision thresholds. The impetus for this recommendation is not without merit, but we believe the conclusion that schools should use different cut scores is unwarranted. The rationale perhaps lies in some schools’ inability to support all students at risk. If teachers in one school screen their first-grade students with DIBELS ORF and find that 75% of students require some intervention, they would not likely have access to the resources to provide those supports. This finding might lead to those educators to consider lowering the decision threshold for

“at risk” status for their students. Such a strategy, however, would mislabel some students as typically achieving when they are not, which could lead to a number of adverse consequences. For example, parents who move their children to a different school might be unpleasantly surprised to find that their “typically reading” child is now at significant risk of reading failure.

Kloo and Zigmond (2008) reported on a school where only 18% of first-grade students had no risk of reading difficulty based on ORF scores. Rather than suggesting a lower standard for risk, they recommended that the school adopt an empirically supported core curriculum, which was not in place, and that teachers teach reading for 90 uninterrupted minutes per day. Kloo and Zigmond also asked schools to introduce small-group instruction. The students in this school improved: in the following year, 45% of first graders had no risk of reading difficulty. Had this school changed their cut scores, they would have reported that some larger proportion of their students were reading at an appropriate level, so administrators may have been less willing to consider a change to their core curriculum. If schools have a strong core curriculum and a tiered support system in place and still find a large proportion of students at risk, it is possible that not all students at risk can receive the interventions they need. We believe it may be more suitable, ethically and politically, to accurately report the proportions of at risk students and the inability to fund interventions for all the students that require help.

## Conclusion

In this chapter, we have summarized the literature on signal detection theory and offered an introduction to procedures for the evaluation of diagnostic or classification systems. In the summary of the section on signal detection theory, we offer some of the key best practices, with references to additional details (e.g., the STARD statement), which will help researchers and consumers of research understand this common approach to the evaluation of screeners and the choice of decision thresholds. Fluency-based screeners were among the first educational measures to offer criterion-referenced cut scores to teachers that predict the future performance of students, and the increased use of the methods will improve the evaluation and selection of screeners and diagnostic tests for literacy, mathematics, student behavior, and other key outcomes in schools.

Fluency in any academic or behavioral skill typically requires fluency with subskills. Fluent readers typically learn first to decode fluently; students fluent at multiplication must first attain fluency with addition. Fluency screeners help teachers identify whether students have achieved normative progress, and with a sufficient set of screeners, teachers can use screeners to determine whether students struggle with subskills. A struggling first-grade reader, for example, may simply need more practice reading connected text, but her challenges might stem from disfluency with certain letter sounds. These two potential diagnoses (among others) suggest different interventions: additional practice reading connected text versus practice with

the sounds for specific subset of letters. Unlike the orchestra conductor, however, who can easily hear an out-of-tune cello, a rushed drumbeat, or a missed note by the pianist, identifying students who have not achieved fluency with blending sounds, single-digit addition, or the doubling rule for spelling (e.g., stop + ing = stopping) is not so easy. Fluency screeners offer teachers a tool to help make such discriminations easier and more objective.

Screeners, however, do not mandate intervention. In schools with limited resources, students who are at risk for reading difficulties yet at the better-performing end of the continuum might not receive immediate intervention. In some cases, teachers may choose to monitor those students more closely to determine if they continue to improve without more intensive services. The identification of specific subskills, however, can also improve the efficiency of instruction, which is critical for resource-strapped districts. As the evaluation of screeners improves and new tests more clearly identify subskills, teachers can better focus instruction for their students. Focused instruction for students, like marking for dancers (Warburton et al., 2013), should reduce the cognitive load and more efficiently lead to fluency gains.

## References

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.
- Bengtsson, S., Nagy, Z., Skare, S., Forsman, L., Forssberg, H., & Ullén, F. (2005). Extensive piano practicing has regionally specific effects on white matter development. *Nature Neuroscience*, 8(9), 1148–1150.
- Brooks, H. E. (2004). Tornado-warning performance in the past and future: A perspective from signal detection theory. *Bulletin of the American Meteorological Society*, 85(6), 837–843.
- Burkel, R. H., Chiou, C.-P., Keyes, T. K., Meeker, W. Q., Rose, J. H., Sturges, D. J., Thompson, R. B., & Tucker, W. (2002). *A methodology for the assessment of the capability of inspection systems for detection of subsurface flaws in aircraft turbine engine components (Final Report, DOT/FAA/AR-01/96)*. Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Research.
- Carran, D. T., & Scott, K. G. (1992). Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics in Early Childhood Special Education*, 12, 196–211.
- Carter, A. S., Briggs-Gowan, M., & Davis, N. O. (2004). Assessment of young children's social emotional development and psychopathology: Recent advances and recommendations for practice. *Journal of Child Psychology and Psychiatry*, 45, 109–134.
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354–380.
- Clarke, B., Baker, S. K., Smolkowski, K., & Chard, D. (2008). An analysis of early numeracy curriculum-based measurement: Examining the role of growth in student outcomes. *Remedial and Special Education*, 29(1), 46–57. doi:10.1177/0741932507309694.
- Connolly, T., Arkes, H. R., & Hammond, K. R. (Eds.). (2000). *Judgment and decision making: An interdisciplinary reader* (2nd ed.). New York: Cambridge University Press.
- Cook, B. G., & Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children*, 79(2), 135–144.

- Dana, J., & Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics*, 29(3), 317–331. doi:10.3102/10769986029003317.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571–582. doi:10.1037/0003-066X.34.7.571.
- Dawes, R. M. (1986). Representative thinking in clinical judgment. *Clinical Psychology Review*, 6, 425–441. doi:10.1016/0272-7358(86)90030-9.
- Dompnier, B., Pansu, P., & Bressoux, P. (2006). An integrative model of scholastic judgments: Pupils' characteristics, class context, halo effect and internal attributions. *European Journal of Psychology of Education*, 21(2), 119–133.
- Engelmann, S., & Carnine, D. (1991). *Theory of instruction: Principles and applications* (Rev. Ed.). Eugene: ADI Press.
- Ericsson, K. A., Krampe, R. T. H., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100, 363–406.
- Ericsson, K. A., Roring, R., & Nandagopal, K. (2007). Giftedness and evidence for reproducibly superior performance: An account based on the expert performance framework. *High Ability Studies*, 18(1), 3–56.
- Fields, R. D. (2005). Myelination: An overlooked mechanism of synaptic plasticity? *The Neuroscientist*, 11(6), 528–531.
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.
- Glover, T. A., Albers, C. A., & Kratochwill, T. R. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, 45, 117–135.
- Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monograph*, 7, No. 2. (Fort Worth, TX: Texas Christian University Press).
- Gredler, G. R. (2000b). Early childhood screening for developmental and educational problems. In B. A. Bracken (Ed.), *The psychoeducational assessment of preschool children* (3rd ed.) (pp. 399–411). Needham Heights, MA: Allyn & Bacon.
- Grove, W. M. (2005). Clinical versus statistical prediction: The contribution of Paul E. Meehl. *Journal of Clinical Psychology*, 61(10), 1233–1243.
- Grove, W. M., & Lloyd, M. (2006). Meehl's Contribution to Clinical Versus Statistical Prediction. *Journal of Abnormal Psychology*, 115(2), 192–194. doi:10.1037/0021-843X.115.2.192.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1), 19.
- Hamilton, C., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review*, 32(2), 228–240.
- Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the dynamic indicators of basic early literacy skills and the comprehensive test of phonological processing. *School Psychology Review*, 32(4), 541–556.
- Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review*, 36(4), 582–600.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Katz, D., & Foxman, B. (1993). How well do prediction equations predict? Using receiver operating characteristic curves and accuracy curves to compare validity and generalizability. *Epidemiology (Cambridge, Mass.)*, 4(4), 319–326.
- Kingslake, B. (1983). The predictive (in)accuracy of on-entry to school screening procedures when used to anticipate learning difficulties. *British Journal of Special Education*, 1, 23–26.
- Kloo, A., & Zigmond, N. (2008). *Implementing progress monitoring in a really low achieving school among very low-skilled teachers*. Paper presented at the 2008 annual Pacific Coast Research Conference.
- Kopiez, R., & Lee, J. I. (2006). Towards a dynamic model of skills involved in sight reading music. *Music Education Research*, 8(1), 97–120.

- Kraemer, H. (1992). *Evaluating medical tests: Objective and quantitative guidelines*. Newbury Park: Sage.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, 38, 404–415.
- Lewinsohn, P. M., Seeley, J. R., Roberts, R. E., & Allen, N. B. (1997). Center for Epidemiological Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. *Psychology and Aging*, 12(2), 277–287.
- Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95(4), 492–527.
- MacGinitie, W., & MacGinitie, R. (2006). *Gates-MacGinitie reading tests* (4th ed.). Iowa City: Houghton Mifflin.
- Malhotra, R., & Indrayan, A. A. (2010). A simple nomogram for sample size for estimating sensitivity and specificity of medical tests. *Indian Journal of Ophthalmology*, 58(6), 519–522.
- Marston, D., Muyskens, P., Lau, M., & Canter, A. (2003). Problem-solving model for decision making with high-incidence disabilities: The Minneapolis experience. *Learning Disabilities Research and Practice*, 18(3), 187–200.
- Martin, S. D., & Shapiro, E. S. (2011). Examining the accuracy of teachers' judgments of DIBELS performance. *Psychology in the Schools*, 48(4), 343–356. doi:10.1002/pits.20558.
- Mason, S. J., & Graham, N. E. (1999). Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting*, 14, 713–725.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.
- Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment*, 50, 370–375.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52(3), 194–216.
- Meisels, S. J. (1987). Uses and abuses of developmental screening and school readiness testing. *Young Children*, 42(4–9), 68–73.
- Nelson, J. M. (2008). Beyond correlational analysis of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS): A classification validity study. *School Psychology Quarterly*, 23(4), 542–552.
- Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, 231, 289–337.
- Pearson Education, Inc. (2007). *Stanford achievement test-10th Edition (SAT10): Normative update*. Upper Saddle River: Author.
- Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford: New York.
- Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *IRE Professional Group on Information Theory PGIT*, 4, 171–212.
- Petscher, Y., Kim, Y.-S., & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention*, 36(3), 158–166.
- Piasta, S. B., Petscher, Y., & Justice, L. M. (2012). How many letters should preschoolers in public programs know? The diagnostic efficiency of various preschool letter-naming benchmarks for predicting first-grade literacy achievement. *Journal of Educational Psychology*, 104(4), 945–958.
- Posner, M. I., DiGirolamo, G. J., & Fernandez-Duque, D. (1997). Brain mechanisms of cognitive skills. *Consciousness and Cognition*, 6(2–3), 267–290.
- Richmond, E. (2012). Different Goals for Students of Different Races? *The Atlantic*.
- Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*. *Law and Human Behavior*, 29(5), 615–620.

- Schatschneider, C., Petscher, Y., & Williams, K. M. (2008). How to evaluate a screening process: The vocabulary of screening and what educators need to know. In L. Justice & C. Vukelich (Eds.), *Achieving excellence in preschool literacy instruction* (pp. 304–316). New York: Guilford Press.
- Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23*, 304–325.
- Smolkowski, K., & Cummings, K. (2014). Evaluation of diagnostic systems: The selection of students at risk for reading difficulties with DIBELS measures (6th edition). Manuscript submitted for publication.
- Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the Classroom Observations of Student-Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly, 27*(2), 316–328. doi:10.1016/j.ecresq.2011.09.004.
- Smolkowski, K., Cummings, K. D., & Baker, D. (2014). Evaluation of diagnostic systems: the selection of English learners at risk for reading difficulties with DIBELS measures (6th edition). Manuscript submitted for publication.
- STARD Statement (2008). *Standards for the Reporting of Diagnostic accuracy studies*. <http://www.stard-statement.org>. Accessed 15 May 2014.
- Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal Of Personality Assessment, 81*(3), 209–219.
- Streiner, D. L., & Cairney, J. (2007). What's under the ROC? an introduction to receiver operating characteristics curves. *Canadian Journal of Psychiatry, 52*(2), 121–128.
- Swets, J. A. (1973). The relative operating characteristic in Psychology. *Science, 182*(4116), 990–1000.
- Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99*(1), 100–117.
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285–1293.
- Swets, J. A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Hillsdale: Lawrence Erlbaum Associates.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000a, October). Better decisions through science. *Scientific American, 283*(4), 82–87.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000b). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1–26.
- Warburton, E. C., Wilson, M., Lynch, M., & Cuykendall, S. (2013). The cognitive benefits of movement reduction: Evidence from dance marking. *Psychological Science*. Advance online publication. doi:10.1177/0956797613478824.
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*, 32–35.
- Zhou, X.-H., McClish, D. K., & Obuchowski, N. A. (2002). *Statistical methods in diagnostic medicine*. New York: Wiley.



## Chapter 9

# Different Approaches to Equating Oral Reading Fluency Passages

Kristi L. Santi, Christopher Barr, Shiva Khalaf and David J. Francis

This chapter examines different solutions to the problem of test equating as it relates to the measurement of oral reading fluency (ORF) using different text probes, a problem that Francis et al. (2008) referred to as “form effects.” The chapter begins with an overview of ORF and curriculum-based measurement (CBM) and a general introduction to the problem of form effects in CBM. Information is provided about form effects and the problems they cause in accurately measuring student progress using CBM as well as about the reasons why equating passages through readability formulas alone is insufficient to ensure form equivalence and a constant measurement scale. Using a middle school data set, the next section of the chapter provides examples of different methods of equating reading probes. Two of these methods, specifically linear equating and equipercentile equating, focus on the equating of raw scores, whereas other methods involve the use of latent variables (LVs) to equate test forms. The latter methods include both linear and nonlinear equating using LVs. This section also includes information on converting unequated raw scores to factor scores on a common scale. The chapter concludes with a discussion of the implications of using equating to improve results when using CBM either to screen for reading problems or risk of reading problems, or to measure student growth over the course of the academic year.

---

K. L. Santi (✉) · S. Khalaf  
College of Education, University of Houston, Houston, TX, USA  
e-mail: klsanti@uh.edu

S. Khalaf  
e-mail: shiva.khalaf@times.uh.edu

C. Barr · D. J. Francis  
Texas Institute for Measurement, Evaluation, and Statistics,  
University of Houston, Houston, TX, USA  
e-mail: Chris.Barr@times.uh.edu

D. J. Francis  
e-mail: dfrancis@uh.edu

© Springer Science+Business Media, LLC 2016  
K. D. Cummings, Y. Petscher (eds.), *The Fluency Construct*,  
DOI 10.1007/978-1-4939-2803-3\_9



## Oral Reading Fluency

Fluency, the ability to read text aloud with speed, accuracy, and prosody, is an important skill in reading development due to its ability to serve as a proxy for comprehension and its consequent value in the identification of children at risk of reading difficulty (National Reading Panel; NRP, 2000; Deno, Fuchs, Marston, & Shinn, 2001; Fuchs, Fuchs, Hosp, & Jenkins, 2001; Snow, Burns, & Griffin, 1998). When placed in the appropriate developmental perspective, ORF is indicative of efficient word-level processing, a rich vocabulary knowledge base, and an understanding of text (Kame'enui & Simmons, 2001). As a result, measures of ORF have widespread utility as means of identifying and monitoring students' progress in overall reading skills (Cummings, Atkins, Allison, & Cole, 2008). Although ORF measures were originally idiographic, and truly based on local curricula, a number of standardized measurement systems with common sets of reading passages have since been developed; for example, Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002a), Texas Primary Reading Inventory (TPRI: TEA, UTHSC, & UH, 2010), Read Well (Sprick, Howard, & Fidanque, 1998), Continuous Monitoring of Early Reading Skills (Mathes, Torgesen, & Herron, 2008), and the Texas Middle School Fluency Assessment (TMSFA; Francis, Barth, Cirino, Reed, & Fletcher, 2010). Attention to reading fluency has increased through the literature base on CBM, which employs standardized oral reading tests that serve as indicators of overall reading achievement (Fuchs et al., 2001).

Students' ORF ability is measured by computing the total number of words read correctly in a fixed unit of time, typically 1 min, on connected text. Typically, ORF scores are then plotted over time to measure students' growth in reading rate. These measures of student growth in fluency (i.e., the level of performance and slope) provide teachers with a means of tracking students' overall reading progress and identifying instructional modifications that might lead to improved learning (Deno, 2003; Fuchs & Stecker, 2003; Shinn, 2002). In order for ORF assessments to accurately measure students' true growth in reading ability, the CBM probes used to measure fluency must meet several criteria. Essentially, these criteria culminate in the individual assessments being reliable and valid. Reliability serves to index the degree to which scores are consistent. One way to think of this consistency is to imagine administering the same CBM fluency assessment to the same individuals on two separate occasions that are relatively close in time (e.g., 30 min to 3 days apart). Probes with reliable scores would yield similar rank orderings of the individuals on the two occasions because one would not expect any true change in fluency, or differences between individuals in the amount of true change between the two occasions of measurement. Consequently, individuals' relative standing on the two measures would be similar, even though individuals' actual scores on the two occasions may be quite different reflecting differences in the difficulty of the passages used on the two separate occasions. Using unequated, but reliable test scores, a teacher could determine

which students are the strongest readers in her class and which are the weakest; however, a teacher interested in tracking student progress, in determining the amount of progress made by individual students, or in ordering the students based on their progress, requires test scores that are equated across text probes in addition to being reliable. When using CBM as a tool for progress monitoring, it is not enough for the rank orderings of students to remain consistent across different probes. The scores themselves must remain consistent. To accurately measure progress, it is essential that differences in test scores from different probes reflect only differences in true ability and differences in the errors in measurement. Having reliable scores ensures that the errors of measurement are small, but does not guarantee that the differences in test scores are due only to differences in ability and differences in error. To ensure that test score differences reflect differences in ability requires a more rigorous process for developing and scoring the passages, and a formal process for equating scores from different probes. In the absence of such a formal equating process, the amount of progress that students make from one time to the next will depend on the order in which the probes were used over time. Only when probes yield equivalent scores can the probes be used over time to accurately convey how much progress individual students are making in reading. A teacher administering a series of equated probes over time can be confident that scores are not going up over time simply because students are reading easier passages, or going down because students are reading more difficult passages. These effects on CBM scores have been removed through the equating process. Any upward or downward movement in the scores is a combination of two factors—changes in ability and error in measurement. The measurement errors can be minimized through the use of reliable scores and by averaging multiple CBM probes at each measurement occasion.

Although we often speak of assessments being valid, validity is not so much an attribute of assessments, but rather an attribute of the inferences that we wish to make based on those assessments (Messick, 1988). One way to think about the validity of inferences from CBM assessments is with respect to inferences about status and growth. If we say that a student is reading at grade level based on a CBM assessment (i.e., we make an inference about status), the accuracy of that statement is a reflection of test validity. If we say that one student is making more progress than another student (i.e., we make an inference about growth), or that a student is making sufficient progress to be on grade level by the end of the year, the accuracy of these statements is a reflection of validity. Of course, the same concerns apply when children's performance is not up to expectations. For students who are not making sufficient progress towards end of the year grade level expectations, this lack of progress must reflect a lack of progress in the student's reading as opposed to a flaw in the test construction. How assessments are developed and the procedures used to administer them are part and parcel to their reliability and validity.

To ensure that CBM assessments measure ORF reliably and that inferences based on CBM ORF assessments are valid, developers need to attend to various aspects of CBM probe development and CBM administration procedures. First,

standardization of test administration procedures (i.e., item sample, duration, administration directions, and scoring) can help to ensure consistency across different administrations; both administrations by the same assessor to different examinees and administrations to the same examinees by different assessors. Second, the procedures for administering the assessment and computing the scores should be efficient and easy for teachers to follow, as teachers are most commonly the individual in schools tasked with assessing students on CBM ORF measures. Third, the CBM ORF probes used to monitor reading progress (i.e., the short texts that students are asked to read) must be grade appropriate and, more importantly, of equivalent difficulty. Finally, generated scores (total words attempted minus errors) should be established on a constant metric and free from measurement artifacts (i.e., practice or form effects) (Shinn, Rosenfield, & Knutson, 1989; Fuchs & Deno, 1991; Martson, 1989). These principles apply whether one is attempting to construct a scale for use within a single grade, or to construct a developmental scale for use across multiple grades. In the case of ORF, the problem of vertical equating (i.e., creating a scale to span multiple years of development in reading fluency) is less challenging than vertically equating tests of other cognitive abilities where the construct being measured may change qualitatively over time.

Research has provided substantial evidence that the first two criteria above can be met with standard approaches to develop and administrate CBM probes to measure ORF, whether based on 1 or 3 min of reading by the student. There is substantial evidence to indicate that ORF procedures can be standardized, they are efficient and easy to use, and can be used effectively by teachers, teacher aides, and researchers. In addition, administration procedures have been standardized across tests to ensure consistent measurement of ORF ability. Further, ORF assessments are clearly both efficient and easy to administer because a child can be assessed, and his or her performance can be scored in less than 5 min (Shinn, 2002). That CBM procedures routinely meet these criteria speaks to the general reliability of CBM assessments. Ample evidence also exists supporting the validity of CBM assessments for predicting status on other reading assessments, such as non-CBM measures of reading fluency and more traditional measures of reading comprehension, and for predicting academic progress (see Deno, 2003; Deno, Fuchs, Marston, & Shin, 2001; Madelaine & Wheldall, 2004).

As alluded previously, one challenge to monitor reading progress over time is the equating of scores that come from different reading probes. Related to this problem of score equating, is the equating of materials used to generate those scores. When texts differ in difficulty, then ORF fluency scores will vary because of these differences in text difficulty. Although these differences may be inconsequential to the relative status of children to one another, they are crucial to inferences about status relative to grade-level benchmarks (Petscher & Kim, 2011) and to differences between students in progress (Francis et al., 2008). In the next section, we review different ways in which CBM developers attempt to control the difficulty of text to ensure comparability of ORF passages. Subsequently, we show that the simple control of ORF passage difficulty is insufficient to ensure the scores generated from those passages reflect a constant scale for the assessment of fluency.

## Form Equivalence as Measured by Readability Formulas

A common method for providing evidence of equivalent passage difficulty is by using readability formulas (Betts, Pickart, & Heistad, 2009). Readability formulas objectively estimate, without measuring characteristics of readers, the difficulty of written material, which is expressed through a numerical score that differs from one formula to the next, both in terms of its construction and in terms of its expression. These scores are often expressed as an estimated grade level (Bailin & Grafstein, 2001) intended to convey the idea that an average reader in that grade should be able to read or “cope” with the text without undue frustration (Begeny & Greene, 2014; Bailin & Grafstein, 2001; Compton, Appleton, & Hosp, 2004). The different readability formulas take into account a combination of factors that contribute to text difficulty, including, but not limited to the percentage or density of high frequency, easy words (e.g., evaluated against a predetermined list of familiar words to most students in a particular grade, or words that frequently occur in text at a given grade level as evidenced by a standard corpus of word frequency), the percentage or density of difficult words (e.g., words that are not included on the list of familiar words, or words that occur infrequently in text at a given grade), the average number of words per sentence, the average number of syllables per word, the number of single-syllable words, or the number of words with multiple syllables (Begeny & Greene, 2014).

Although the first instance of concern about difficulty levels of written material goes back to 900 AD when the Talmudists counted the number of occurrences of words and individual ideas in their scrolls, a scientific approach to text difficulty and readability originated in the 1920s (Tekfi, 1987; Klare & Buck, 1954). Throughout the 20th century, several formulas have been developed and used for different purposes, such as, determining the readability of government documents, newspaper articles, schoolbooks, and medical documents (Begeny & Greene, 2014; Bailin & Grafstein, 2001). For example, the Forcast formula, which is calculated based on the number of one-syllable words per 100 words, was developed to evaluate exams and entrance forms for the US Army (Sticht, 1973). Table 9.1 provides a display of some of the more commonly used readability formulas utilized in education today along with the corresponding formulas. Of the formulas for measuring text difficulty listed in Table 9.1, only the Lexile framework was developed as a metric that could be applied to both readers and texts in order to place both on a common, interval scale; that is to say, a scale where numerical differences of a given magnitude mean the same thing no matter where they appear in the scale. The Lexile framework is based on a theoretical formulation of the factors that affect the comprehensibility of text and is backed by empirical research that supports the theoretical formulation. The objective behind the framework is to place readers and texts on the same scale such that the score assigned to a student indicates the level of text that the student can read with a specified level of understanding. Therefore, by analyzing a student’s performance on calibrated reading tasks, she/he is then located on the Lexile scale at the point where she/he is predicted to achieve

**Table 9.1** Summary of readability formula

Name	Formula
Dale-Chall Dale & Chall, 1948	$\text{Grade} = (0.1579 \times \text{percent unfamiliar words}) + (0.0496 \times \text{word/sentence}) + 3.6365$
Flesch-Kincaid Flesch, 1948	$\text{Grade} = 0.39 (\text{average words/sentence}) + 11.8 (\text{average syllables/word}) - 15.59$
FOG Gunning, 1952	$\text{Grade} = 0.4 [(\text{average words/sentence}) + (\text{percent of hard words})]$
Forcast Sticht, 1973	$\text{Grade} = 20 - (\# \text{ single-syllable words}/10)$
Fry Fry, 1968	Grade = graph average (#sentence, #syllables) from three 100-word passages
Lexile Stenner et al., 2007	Based on word frequency and sentence length (rounded to the 10L)
SMOG McLaughlin, 1969	$\text{Grade} = (0.121 \times \text{word/sentence}) + (0.082 \times \text{percent unfamiliar words}) + 0.659$
Spache Spache, 1953	$\text{Grade Level} = (0.141 \times \text{ASL}) + (0.086 \times \text{PDW}) + 0.839$ (AWL = average sentence length and PDW = percent of difficult words)

75% comprehension of the text. That is, a student with a Lexile measure of 500L is estimated to correctly answer 75% of Lexile items sampled from a text measuring 500L (Stenner, Burdick, Sanford, & Burdick, 2007). Thus, the more difficult the passage, the higher the Lexile measure for that text, and the more skilled the student must be to comprehend the text.

Although readability statistics are widely used by teachers as tools for selecting reading materials that are matched to students' reading proficiency levels (Hiebert, 2002) and for selecting passages that are of comparable difficulty to one another, researchers have questioned the use of these formulas for over 50 years (e.g., Swanson & Fox, 1953). A number of studies have examined the relation between ORF and readability and the extent to which readability accounts for variability in ORF across CBM probes. These studies have generally found weak correlations between readability levels and ORF for both struggling readers and typically developing readers (Compton et al., 2004; Powell-Smith & Bradley-Klug, 2001). Other studies have evaluated the validity of readability formulas as predictors of student's ORF rates and found weak relations between ORF rates and passage difficulty (Ardoin, Suldo, Witt, Aldrich, & McDonald, 2005). Finally, researchers have also examined variability in estimates of readability when different formulas are applied to the same CBM probes. For example, Good and Kaminski (2002b) developed passages that were targeted at the end of the year or the beginning of the next grade year using the Spache readability index. However, using other readability indices, the same passages showed a markedly different range of readability estimates depending on the formula used to evaluate them. For example, passages with Spache indices ranging from 2.4 to 2.7 were found to range from 4.3 to 8.0 on the FOG, 3.0–6.6 on the Fry, 6.8–9.5 on the Forcast, and 2.2–5.3 on the Flesch (see Table 9.4, pg. 7, Good & Kaminski, 2002b). At first glance, this inconsistency across measures may seem

surprising; however, on closer examination, it would seem to be expected. On the one hand, readability formulas are designed to estimate the difficulty that students may encounter in understanding a text, and ORF probes are used to measure fluency as a proxy for comprehension. ORF probes do not measure comprehension directly, but only indirectly through the link between fluency and comprehension. Second, the link between readability and comprehension is also imperfect in so far as measures of readability are imperfectly related to students' comprehension, although readability and comprehension are more strongly correlated than readability and fluency.

Recently the notion of readability has evolved into a more general notion of text difficulty, with significant contributions from researchers interested in text and discourse and cognitive models of reading comprehension. These approaches attempt to go beyond the features of words and sentence composition to examine more complex features of text. Several of the more comprehensive approaches to examining text difficulty were recently reviewed by Nelson, Perfetti, Liben, and Liben (2012) in a report for the Council of Chief State School Officers. These researchers found variability across the measures, but nevertheless found strong correlations (generally in the range from .5 to .8) with reference indicators of text difficulty. Specifically, the study involved five reference sets of text that had been ordered a priori according to difficulty. For example, one set of texts was comprehension passages from the Gates-MacGinitie Reading Test, and a second was comprehension passages from the Stanford Achievement Test a third set was passages from state standardized reading assessments. These passages can be ordered in difficulty a priori based on the grade level in which they are used. More importantly, the researchers found strong relations between text difficulty and measures of students' reading ability as measured by the average item difficulty for questions associated with specific texts. These correlations tended to be higher (generally between .7 and .8 with one exception) than the correlations between the text difficulty measures and grade level. The researchers further concluded that the more comprehensive measures, such as CohMetrix (Graesser, McNamara, & Kulokowich, 2011), better explained text difficulty than measures that were limited to word frequency and sentence length.

## **Beyond Form Equivalence to Score Equivalence**

Even though readability and text difficulty measures provide useful information about the relative difficulty of texts, the measurement of growth requires a more stringent equivalence criterion among ORF CBM probes than is possible through typical form development standards and readability formulas. A large number of research studies have shown that estimates of text difficulty based on readability formulas are not adequate for ensuring that reading materials are equivalent (see Begeny & Greene, 2014). For example, Betts et al. (2009) found that although readability statistics were able to identify course differences in difficulty levels



between different grades, they were not able to identify fine differences *within* a grade level. Equivalence of test scores is difficult to achieve through design specifications alone. This problem is not unique to the development of CBM ORF probes, but until recently had been largely ignored in the CBM literature. Test developers routinely go through processes for equating scores from different forms of the same test so that students are neither advantaged nor disadvantaged by the specific test form that they happen to take on a particular day. The same has not been routinely done in the development of CBM ORF probes. Using different probes from a published CBM ORF assessment, Francis et al. (2008) found substantial differences in estimates of ORF scores depending on which probes students were asked to read. More importantly, they showed that these passage effects also changed the shape of growth trajectories and affected estimates of linear growth rates. At the same time, Francis et al. (2008) demonstrated how the same procedures used to equate different test forms for traditional assessments could be used to remove form differences in CBM ORF probes. By using explicit equating procedures with CBM probes to remove passage effects, Francis et al. (2008) showed that explicit equating is essential to the development of equivalent forms, which in turn is essential to make valid inferences about growth. Although scores from different test forms were highly correlated, meaning the scores rank-order students in approximately the same order across forms, variability in difficulty affected students' absolute levels of performance in predictable ways and consequently affected estimates of student growth.

When measuring skills like comprehension or reading fluency, we cannot ignore the stimulus materials and their influence on our estimate of true scores. Many elements of stimulus construction can be averaged across administrations to diminish their influence on true score estimation. For example, we can use both narrative and expository passages in a measure of reading comprehension, and we can select passages from a wide variety of subject matter areas so that individuals are not advantaged or disadvantaged by the topical focus or genre of the reading passages. It would be theoretically possible to average out the effects of passage difficulty in the assessment of ORF if one were willing to have students read multiple passages each time they were tested and were willing to average the different ORF scores for the student. But interest in efficiency has dominated assessment procedures in CBM, so much so that even when multiple assessments are used, it is common to limit the number of passages to three and to take the median ORF score as the most stable one. Although the median of three measures is less variable than any one measure, it is more variable than the average of those measures. Most importantly, the median will not have the same expected value as the mean if the three measures are not equally difficult. The expected value of the median of three measures is the population median of the three measures, whereas the expected value of the mean is the population average of the three measures. However, when the three passages are not equally difficult, the mean and median are not equally effective in averaging out the effects of text difficulty. (For a detailed examination of using the median to estimate ability in CBM procedures, see Barth et al., 2012; for a more detailed comparison of the mean and median for estimating ability, see Petscher and Kim, 2011.)



The difference between the mean and median for estimating ORF ability could be diminished and the effects of passage difficulty could be eliminated, if the median and mean were computed on equated scores, rather than raw scores. What is needed is not for students to read more passages, but to equate the scores from passages that differ in difficulty so that all scores are placed on a common scale. In what follows, we will use data from a project with middle school students to demonstrate several practical ways that test developers can mitigate the effects of text difficulty on the estimation of students' ORF ability. We contend that it is not possible, nor is it desirable, to have passages that exist at a single level of text difficulty. Rather, passages should sample from the range of text difficulty that students are likely to encounter in their grade. What teachers need is a mechanism for removing the effects of text difficulty on a student's estimated ORF ability so that, from the standpoint of the student and teacher, it is irrelevant which passage(s) the student reads on any particular day and what results from the assessment is an accurate reflection of the student's ability to read grade-level text. This mechanism for placing all passage scores on the same scale is known as *equating*. Through a formal equating process test developers can provide teachers and students with scores on a constant scale that remove the effects of varying levels of text difficulty on the estimation of students' ORF ability. We begin the remainder of the chapter by first introducing the dataset on which the examples in this chapter are based followed by a description of several different approaches to equating CBM probes along with applications of these approaches to the sample dataset.

## Equating CBM Probes for Middle School Students

### *The Texas Middle School Fluency Assessment*

The TMSFA is a reading instrument designed to measure growth in reading fluency for students in grades 6, 7, and 8. The TMSFA measures students' ability to both recognize words by sight in the absence of context and to identify words while reading connected text. The reading fluency skills measured by the TMSFA are essential to the development of overall reading ability. As a result, the TMSFA has been developed to help identify why certain middle school students are lagging behind their peers. Moreover, like all ORF measures, the TMSFA permits frequent assessments either by a classroom teacher, diagnostician, or other testing professional (Fletcher, Lyon, Fuchs, & Barnes, 2007; Snow, Burns, & Griffin, 1998) in order to evaluate student progress.

The TMSFA was developed in the summer of 2006, piloted in August, 2006, and implemented in the early fall of 2006, with additional testing sessions taking place in December, 2006, and January, February, and April, 2007. A total of 1867 students in grades 6–8 participated in the validation studies for the TMSFA, 733 students in grade 6, 450 students in grade 7, and 684 students in grade 8. Across grades, there

were 754 typical readers; this subgroup comprised 40% of the total sample. The remaining 1113 students were identified as struggling readers. Gender and ethnicity information was also obtained. The ethnic breakdown was as follows: 42% of the sample was African American, 36% Hispanic, 19% Caucasian, and 3% Asian. Furthermore, both genders were almost equally represented (49.5% female). Lunch status was not available for 60 students, but of the remainder, 68% were eligible for free or reduced lunch. Thus, the sample was ethnically and economically diverse (Francis et al., 2010).

The TMSFA consists of two subtests: (a) the word reading fluency measure and (b) the passage reading fluency measure. In addition to receiving the TMSFA measures, students participating in the project were also assessed on both subtests of the Tests of Word Reading Efficiency (TOWRE; Torgesen et al., 2001; viz., phonemic decoding efficiency and sight word efficiency), the Test of Sentence Reading Efficiency (TOSRE; Wagner, Torgesen, Rashotte, Pearson, 2010), and the Test of Silent Contextual Reading Fluency (TOSCRF; Hammill et al., 2006). Table 9.2 provides a summary of all measures administered to students and the order in which they were administered. Because the remainder of the chapter is focused on the equating of the CBM probes for the passage reading fluency measure, we will not say more about the other component of the TMSFA or the validating battery of tests except for three measures used in the LV equating. Interested readers should consult Francis et al. (2010) or contact the chapter authors for additional information.

**Table 9.2** Fluency passage ordering within and across grades

Order read						
Grade	Order	First	Second	Third	Fourth	Fifth
6	A	1	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
7	A	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
8	A	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	7
6	B	8	9	10	11	12
7	B	9	10	11	12	13
8	B	10	11	12	13	14
6	C	15	16	17	18	19
7	C	16	17	18	19	20
8	C	17	18	19	20	21
6	D	22	23	24	25	26
7	D	23	24	25	26	27
8	D	24	25	26	27	28
6	E	29	30	31	32	33
7	E	30	31	32	33	34
8	E	31	32	33	34	35

*Note:* The grey shaded boxes with the bold number indicate the five stories across all grades used in the analysis. Story 2—*Koalas*; Story 3—*Suni*; Story 4—*An Unusual Job*; Story 5—*A Wonderful Friendship*; and Story 6—*The King's Gold*. The remaining stories may be found in the TMSFA.

### ***The Passage Reading Fluency Measure***

The passage reading fluency measure consists of 35 passages that were developed in both narrative and expository text structure. All passages averaged approximately 500 words each. Passages ranged in difficulty from approximately 350 Lexiles to approximately 1400 Lexiles (for more on the Lexile metric see Stenner et al., 2007). Within grades, students were randomly assigned to one of five groups, with each group reading five passages. Within a given grade, different groups read different passages, but across grades there was some overlap in the passages read as shown in Table 9.2. Table 9.2 displays the passages read by the students in each group in each grade and the order in which the passages were read. The group designation in Table 9.2 signifies the grade and the randomly assigned group within grade (A–E). The table makes clear the overlap in stories between students in different grades. For example, in group A, students in grade 6 read stories 1–5, students in grade 7 read stories 2–6, and students in grade 8 read stories 3–7. As can be seen in the table, within a given group set (A–E) there are four passages that overlap between 6th and 7th graders, three passages that overlap between 6th and 8th graders, and four passages that overlap between 7th and 8th graders. The procedures for the study required that students read five stories for 1 min each within a single sitting and then answer approximately eight explicit and inferential questions about the story. ORF scores were recorded individually as the number of words read correctly per minute of reading (words correct per minute (WCPM)). Although students read the first two stories in their entirety, the analyses are based on the WCPM score from the first minute of reading.

### ***Test of Word Reading Efficiency (TOWRE; Torgesen et al., 2001)***

TOWRE is an individually administered test of speeded reading of single words. This measure consists of two subtests: sight word efficiency and phonemic decoding efficiency. The sight word efficiency task requires students to read as many printed *real words* as they can within 45 s. The phonemic decoding task requires students to read as many pronounceable *pseudo words* as they can within 45 s. The internal consistency reliability for the sight word efficiency and the phonemic decoding efficiency tasks is .93 and .94, respectively.

### ***Test of Sentence Reading Efficiency and Comprehension (TOSRE; Wagner et al., 2010)***

TOSRE is a 3-min, group-based assessment of reading fluency and comprehension. Students are presented with a series of short sentences and are required to read each sentence to themselves and then mark on the page next to the sentence if the

statement read was true or false. The raw score is the number of correct minus the number incorrect. The test is both reliable and valid for the age range of interest.

## Linear Equating

Linear equating is the process of placing observed scores on a common metric based on means and standard deviations (Holland & Rubin, 1982). In the case of ORF, linear equating removes mean and variance differences in WCPM across probes. Linear equating defines passages as being equivalent through the following equation:  $\frac{p1i - \mu(P1)}{\sigma(P1)} = \frac{p2i - \mu(P2)}{\sigma(P2)}$ , where  $p1i$  and  $p2i$  are the WCPM scores for person  $i$  on passage 1 and 2, respectively, and  $\mu(Pk)$  and  $\sigma(Pk)$  designate the mean and standard deviation, respectively, for passage  $k$ . This formula stipulates that a given person's score ( $p1i$ ,  $p2i$ ) would be the same on two different probes if we expressed each score as a standardized deviation from the mean ( $\mu(P1)$ ,  $\mu(P2)$ ) for that probe. Linear equating assumes that scores can be equated using the same function at all points of the WCPM score distribution. Put another way, this formula assumes that the score distributions for two probes have the same shape; the distributions simply differ in their means ( $\mu(P1)$ ,  $\mu(P2)$ ) and standard deviations ( $\sigma(P1)$ ,  $\sigma(P2)$ ). By using linear equating for all test forms, we assume that students' true WCPM scores differ between any two CBM probes by a constant difference multiplied by a scaling factor (i.e., the standard deviation). Any other difference between observed WCPM scores for the same student on two different probes is due to error. Error can occur in two forms, random and systematic (Kolen & Brennan, 1995). Random error occurs because the equating relations were created using samples drawn from the population and not the entire population. Random error decreases as sample size increases. Systematic error can come from several sources, including bias in the method of estimation, nonrandom samples, violations to statistical assumptions, etc. Sample size has no impact on systematic error. As a general rule, standard errors of equating should be about .1 at the mean, and can increase to 1 at the Mean  $\pm$  2SD (Kolen & Brennan, 2004).

### *Linear Equating Empirical Example*

In this section, we present an empirical example for linear equating from the TMSFA. Specifically, we will present the linear equating process and results for five stories: *Koalas*, *Suni*, *An Unusual Job*, *A Wonderful Friendship*, and *The King's Gold*. These stories are the five stories that were read by students in grade 7, group A, but four of these five stories were also read by grade 6, group A, and grade 8, group A. Recall from the previous design discussion that there are 15 groups across three grades and students in the same grade in each group read the

**Table 9.3** Descriptive statistics for grade 7 group A ORF passages

Variable	<i>N</i>	Mean	Standard deviation	Minimum	Maximum
Koalas	230	117.02	37.73	3	217
Suni	367	132.05	36.73	26	239
An Unusual Job	367	119.93	37.96	22	214
A Wonderful Friendship	367	124.14	35.24	29	208
The King's Gold	222	129.06	39.77	33	245

same five passages. Across the three grades, each group (A, B, C, D, and E) was assigned seven passages with the sixth graders reading the first to fifth passage, seventh graders read the second to sixth passage, and eighth graders read the third through seventh passage in a set of seven passages. For the current empirical example, we are omitting the first and seventh passages and only focusing on the five passages in the set with some overlap across the three grades. The descriptive statistics for the raw WCPM scores prior to equating can be seen in Table 9.3. As seen in the table, *Koalas* has the lowest observed mean, which is to be expected because students in lower grades would be expected to read slower than students in upper grades and no eighth graders read this probe. Yet the passage *The King's Gold* did not have the highest mean, even though this passage was read by only seventh and eighth grader students, who would be expected to read the fastest. Instead, *Suni* a passage read by all three grades had the highest mean, though it was quite comparable to *The King's Gold* and only differed by  $d = .07$ . Of the remaining passages read by all students, *An Unusual Job* and *A Wonderful Friendship* had observed means less than both *Suni* and *The King's Gold*.

When conducting equating, the typical approach is to generate a referent value that allows for the comparison of observed scores on multiple forms. Three examples of referent values are: (a) equating to arbitrary values on an arbitrary scale, such as equating WCPM to *t* scores (i.e., Mean = 50, sd = 1.0), (b) equating back to the simplest passage values, or (c) equating to composite WCPM values (i.e., an average WCPM computed across multiple probes). These referent values can then be associated with specific WCPM values for the fluency passages and allow instructors to obtain the referent score for any given student's WCPM on any given fluency passage. Thus, WCPM scores can be directly compared across multiple passages. In the current example, we chose to use composite WCPM values. For our referent WCPM scores, we computed the mean and standard deviation for scores across the three passages that were read by students in all three grades: *Suni*, *An Unusual Job*, and *A Wonderful Friendship*. In this way, we used the maximum amount of information from a common set of students as our referent. Table 9.4 displays the unadjusted values of the five passages and the referent value of these passages across the range of scores from  $-2$  standard deviations to  $+2$  standard deviations. The table allows for a direct comparison of WCPM

**Table 9.4** Linear equating of grade 7 group A ORF passages

Reference value	Reference WCPM	Koalas	Suni	An Unusual Job	A Wonderful Friendship	The King's Gold
+2 SD	198.66	192.48	205.51	195.85	194.62	208.6
+1.5 SD	180.34	173.615	187.145	176.87	177	188.715
+1 SD	162.02	154.75	168.78	157.89	159.38	168.83
+0.5 SD	143.70	135.885	150.415	138.91	141.76	148.945
Mean	125.37	117.02	132.05	119.93	124.14	129.06
-0.5 SD	107.05	98.155	113.685	100.95	106.52	109.175
-1 SD	88.73	79.29	95.32	81.97	88.9	89.29
-1.5 SD	70.41	60.425	76.955	62.99	71.28	69.405
-2 SD	52.09	41.56	58.59	44.01	53.66	49.52

*WCPM* words correct per minute

values from different passages back to the common referent. For example, reading 187 WCPM on *Suni* or 177 WCPM on *A Wonderful Friendship* both trace back to a referent value of 180 WCPM. Thus, a student who reads 187 WCPM on *Suni* would have the same expected fluency ability as a student who reads 177 WCPM on *A Wonderful Friendship*.

### Equipercntile Equating

As mentioned, linear equating assumes that the scores being equated are normally distributed and thus all relevant information about distribution differences are contained in the means and standard deviations of the distributions. Problems can arise with linear equating when WCPM score distributions are non-normal, or differently shaped across passages. Equipercntile equating can provide more accurate equated scores, particularly when scores are either very low or very high relative to the mean, but requires large sample sizes because of the reliance on accurate estimation of order statistics. Linear equating is, in fact, considered to be an approximation of equipercntile equating (Hambleton, Swaminathan, & Rogers, 1991). Equipercntile equating does not rely on passage means or standard deviations to equate passages. Instead, equipercntile equating assumes that the passages are equivalent based on percentile ranks of observed WCPM scores. This reliance on percentile rankings allows relations between observed WCPM scores and true scores to be nonlinear. If it is the case that curvilinearity exists in the relation of observed and true scores, equipercntile equating is the only observed score equating method that will account for this curvilinearity. In other contexts, such as tests that yield a number correct as the observed score, equipercntile equating is also better at handling passages when the differences in the difficulties of the forms are large, or

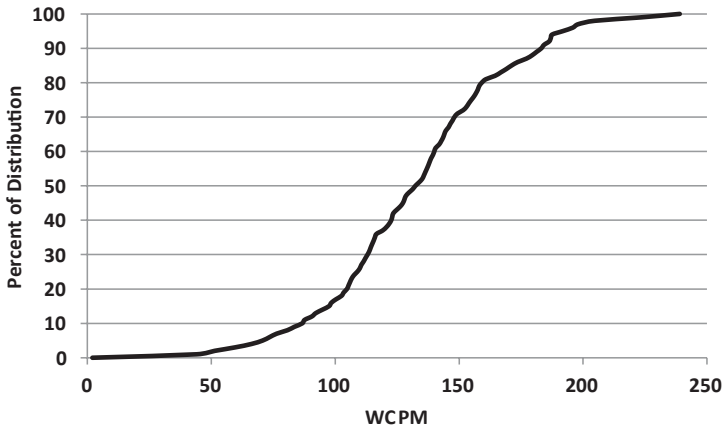
when there is differential strength in the students' abilities. But because equipercentile equating involves the estimation of more parameters than linear equating, the sample size requirements are greater (Kolen & Brennan, 2004).

Generally speaking, equipercentile equating requires significantly larger samples than linear equating. However, the specific sample size required to achieve the same precision with linear and equipercentile equating will depend on many factors, including the expected shape of the distribution and the desired precision in equating at specific points on the score distribution. Kolen and Brennan (2004), in their excellent text on test equating, give an example where the desired precision is .1 raw score standard deviations between  $-2$  and  $+2$  standard deviations of the mean, and the distribution of observed scores is normal. In this particular instance, using a random equating design (i.e., subjects randomly assigned to one test or the other), linear equating would require a sample of 400 students per form, whereas equipercentile equating would require a sample size of 1500 students per form, or roughly four times as many students to achieve the same precision.

In some ways, for CBM equating, the prior example might be considered a best case scenario as raw CBM scores may be non-normal, and interest in CBM is often in scores at the low end of the distribution (i.e., scores well beyond  $-1.5$  standard deviation units below the mean). Of course, if the focus is on low-performing students, it is possible to concentrate sampling on those students and to develop equating rules for use of CBM assessments with at-risk students. That is, it is not necessary to use a sample of 10,000 to get 200 students below the 20th percentile. Rather, sampling can be concentrated on those students for whom the test is to be used for progress monitoring. It is also important to note that, within the subgroup of at-risk students, CBM scores may be more normally distributed, and thus, linearly related across passages, making it possible to use linear equating effectively within this population of students. Of course, if the test is to be used with all students, then a restricted sampling strategy would not lead to the desired equating for all students, regardless of whether linear or equipercentile equating were used. In this instance, an alternative is to oversample at-risk students, providing increased precision in the lower tail of the distribution while preserving the ability to equate scores throughout the range of possible CBM scores and also keeping sample size requirements more manageable.

In equipercentile equating if all possible observed WCPM raw scores do not occur in a given sample, interpolation methods would be required to obtain an equipercentile rank for any unobserved WCPM values. Additionally, at points of the distribution where there is limited information (e.g., fewer scores in the tails of the distribution) the reliability of the estimates will be lower. Oversampling at-risk and high-ability students can reduce the impact of these potential problems and lower the overall sample size requirements to obtain the same degree of precision in the tails of the distribution. Finally, it must be kept in mind that when distributions are similarly shaped, linear equating can be more accurate than equipercentile equating.





**Fig. 9.1** Equipercentile equating plot for *Suni*. The figure depicts the curvilinear nature of the data as it approaches the tails of the distribution

### *Equipercentile Equating Empirical Example*

In this section, we present the empirical example for equipercentile equating. Figure 9.1 presents real data from for the percentiles of *Suni*. As seen in the figure, the relation between the percentiles and WCPM is fairly linear between the 20th and 80th percentiles but becomes curvilinear beyond that. A linear equating framework would likely overestimate performance in the bottom tail and underestimate performance in the top tail. Like the linear equating example above, we will present the equating process and results for five stories: *Koalas*, *Suni*, *An Unusual Job*, *A Wonderful Friendship*, and *The King's Gold*. In conducting equipercentile equating, the cumulative percentage of all scores below a target score are added to  $\frac{1}{2}$  of the percentage at the target score. For example, if 50% of students read less than 100 WCPM and 3% of students read exactly 100 WCPM, then the equipercentile value for 100 is 51.5% ( $50\% + (.5 \times 3\%)$ ). This addition of  $\frac{1}{2}$  of the percentage of examinees at the target value to the cumulative percent essentially treats the value of 100 as representing the interval of values from 99.5 to 100.49. By adding  $\frac{1}{2}$  of the percentage, the examinees at the value 100 are being spread across the interval from 99.5 to 100.49 so that they contribute equally to the cumulative percentage below the target value and the percentage below the next higher value. Although this  $\frac{1}{2}$  percent below and above a specific WCPM seems trivial, it underscores the continuous nature of the underlying fluency construct, in that an observed score may be an integer but its true value can range from  $\pm .5$  of the WCPM. For a complete explanation of this reasoning, see Blommers and Forsyth (1977).

Through equipercentile equating, any WCPM score can be given an equal rank in cumulative percentages (i.e., an equipercentile rank) so long as data are

available about the percentages of students at or below specific WCPM scores from a sufficient number of students. This equating process requires normative data and must be carried out by test developers or researchers with expertise in measurement and familiarity with CBM probes. Test equating is not the responsibility of test users, such as classroom instructors, but of test developers. Typically, the test developer constructs look-up tables for each passage and test users convert obtained WCPM scores during testing to equated WCPM scores through the look-up tables or through software that carries out the table look-up in the background. The look-up table converts the observed WCPM raw score from a specific probe to a percentile rank and then to a scaled score that is equivalent across different test probes. The scale is constructed to have a desired mean and standard deviation. The values for the mean and standard deviation of the scaled score are arbitrary, but could be chosen to reflect the distribution of ORF scores at a specific grade level, or on the distribution for a particular probe used at that grade. Although other scale choices exist, basing the mean and standard deviation of the scaled score on a particular test form, or the average across all forms at a grade are consistent with the general philosophy behind CBM assessment. Regardless of the choice of scale by the test developer, the look-up tables allow classroom instructors to easily and rapidly determine the scaled score value for a given WCPM score from any CBM ORF probe. By using these equated scaled scores, teachers can easily monitor fluency progress on a common metric across passages without concern that scores will fluctuate from one period to the next because of changes in the CBM probe. Although scores will still fluctuate across assessment intervals, one considerable source of fluctuation in the scores has been controlled resulting in scores that are more comparable over time.

We have constructed these look-up tables and present them in Table 9.5 for the five passages read by grade 7, group A students. As seen in the table, a student reading 120 WCPM on *Koalas* scored at roughly the 48th percentile for all students reading that probe. Based on the equating tables, this same student would be expected to read 129–130 WCPM, 118–119 WCPM, 124–125 WCPM, and 123–124 WCPM, on passages *Suni*, *An Unusual Job*, *A Wonderful Friendship*, and *The King's Gold*, respectively. In the construction of the table, we used all available data—grade 6 data on *Koalas*, *Suni*, *An Unusual Job*, and *A Wonderful Friendship*, grade 7 data on *Koalas*, *Suni*, *An Unusual Job*, *A Wonderful Friendship*, and *The King's Gold*, and grade 8 data on *An Unusual Job*, *A Wonderful Friendship*, and *The King's Gold*. The use of all data is preferable because form effects on passages are a characteristic of the passages themselves and not of samples reading the passages. Although it is possible to develop equating tables that are grade specific, we would argue that it is more appropriate to have a scale that can reflect the development of ORF across the entire developmental range where the particular CBM probe has been targeted for use. Using all available data allows for more reliable estimation of the equipercntile rank for values of WCPM across the full range of the underlying fluency ability, that is, over the developmental range of ability for which the probe will be used.

It is important to distinguish equipercntile ranks, which are associated with the raw score distributions for specific ORF passages, from norm-referenced

**Table 9.5** Equipercntile equating of grade 7 group A ORF passages

Equipercntile rank	Koalas	Suni	An Unusual Job	A Wonderful Friendship	The King's Gold
< 1	LE 2	LE 25	LE 21	LE 28	LE 32
1	19.30	44.17	25.06	37.17	38.22
2	35.10	51.34	33.84	46.17	46.44
3	37.40	59.51	43.01	52.51	56.16
4	49.20	66.18	47.39	55.18	60.94
5	58.00	70.68	60.68	65.85	65.60
6	59.90	73.52	63.51	71.51	67.82
7	61.60	76.35	65.56	73.06	73.04
8	63.70	80.68	69.86	74.29	74.69
9	64.85	83.53	72.67	75.52	75.25
10	66.50	86.68	73.28	77.35	77.70
11	67.27	87.62	74.29	79.19	80.21
12	68.03	90.54	75.51	80.26	83.14
13	68.95	92.18	76.24	81.40	84.62
14	70.60	94.88	77.88	84.19	87.04
15	72.00	97.51	79.53	86.85	88.65
16	73.90	98.43	80.93	88.36	90.52
17	76.05	100.30	82.20	89.63	92.44
18	77.90	102.52	84.03	90.85	92.99
19	79.85	103.43	85.41	92.37	94.73
20	82.50	104.73	86.70	93.90	96.20
21	85.58	105.35	88.19	95.27	97.31
22	86.15	105.96	90.08	96.75	98.42
23	87.90	106.60	91.30	97.97	100.52
24	91.30	107.53	94.19	99.19	101.26
25	91.88	108.88	95.42	100.19	102.00
26	93.10	109.90	96.57	100.98	102.86
27	94.55	110.52	97.18	101.72	103.97
28	96.90	111.44	97.94	102.45	105.08
29	98.73	112.19	100.64	103.72	106.46
30	99.50	113.03	102.05	105.53	108.10
31	100.27	113.75	103.19	106.89	108.86
32	101.70	114.28	104.55	107.91	109.31
33	102.47	114.85	104.96	108.56	110.57
34	103.23	115.46	105.36	109.95	111.12
35	104.00	115.99	106.71	111.73	112.70
36	104.90	116.62	107.23	112.92	115.73
37	106.60	119.10	107.86	114.29	116.29
38	108.20	120.65	108.62	115.19	117.77
39	109.85	121.78	109.57	116.54	118.22
40	111.00	122.59	111.10	117.90	118.70
41	112.15	123.00	111.85	118.91	119.26
42	113.80	123.40	112.38	119.55	119.75

**Table 9.5** (continued)

Equipercntile rank	Koalas	Suni	An Unusual Job	A Wonderful Friendship	The King's Gold
43	115.95	124.77	113.44	120.77	120.19
44	117.10	126.24	114.20	121.99	120.67
45	117.88	127.29	115.54	122.93	121.23
46	118.45	127.90	116.46	123.62	121.87
47	119.20	128.43	117.80	124.14	122.61
48	120.20	129.66	118.66	124.69	123.35
49	121.35	131.21	119.97	125.31	123.86
50	122.00	132.33	120.63	126.33	124.30
51	122.55	133.79	121.53	127.29	126.61
52	122.93	135.07	122.27	128.21	127.94
53	123.32	135.75	122.92	129.34	129.33
54	123.74	136.36	123.55	130.14	130.44
55	124.20	136.98	124.46	130.87	131.20
56	125.30	137.57	125.38	131.57	132.66
57	126.20	138.10	126.30	132.02	134.27
58	126.97	138.67	127.22	132.48	135.88
59	127.68	139.41	128.13	133.21	136.99
60	128.25	139.96	129.05	134.55	137.90
61	128.93	140.48	130.44	135.47	138.71
62	129.60	141.89	131.39	136.64	139.82
63	129.98	142.74	132.04	137.71	140.67
64	130.37	143.48	132.60	139.22	141.12
65	131.67	144.01	133.01	140.35	141.58
66	132.43	144.57	133.41	141.31	142.13
67	132.92	145.68	134.97	141.98	142.87
68	133.38	146.41	136.28	142.57	143.98
69	133.84	147.31	137.15	143.03	145.59
70	134.30	148.08	138.13	143.49	146.90
71	135.80	149.28	139.14	143.95	148.81
72	138.30	151.62	139.87	144.41	150.92
73	140.40	152.98	140.49	145.08	153.56
74	142.10	153.90	141.10	145.76	156.78
75	143.00	154.92	141.75	146.38	157.87
76	143.90	155.98	142.48	146.99	158.43
77	146.02	156.82	144.29	147.80	159.47
78	146.48	157.54	146.59	149.25	160.58
79	146.94	158.06	147.69	150.23	161.69
80	148.50	159.10	148.42	151.87	162.80
81	149.80	160.77	151.13	154.64	163.77
82	151.53	164.47	156.19	156.15	164.54
83	152.95	166.80	156.92	157.15	166.76
84	154.60	169.07	159.97	158.26	168.98
85	156.50	171.15	160.49	159.24	171.85

**Table 9.5** (continued)

Equipercntile rank	Koalas	Suni	An Unusual Job	A Wonderful Friendship	The King's Gold
86	158.90	173.62	161.81	162.23	172.96
87	160.55	177.14	163.93	162.76	174.07
88	161.90	179.48	166.98	164.15	177.12
89	164.70	181.31	170.04	165.16	177.86
90	168.00	183.15	171.27	166.27	182.27
91	170.80	184.24	172.98	167.99	183.52
92	173.03	186.32	174.38	170.32	186.12
93	174.95	186.97	176.81	173.65	189.23
94	175.80	187.50	179.99	176.49	191.18
95	176.38	191.82	185.15	182.32	195.40
96	177.40	195.82	189.82	185.27	201.06
97	179.78	198.00	193.49	189.99	205.84
98	180.35	204.66	197.39	198.66	214.56
99	182.20	223.33	205.16	202.66	226.28
>99	GE 217	GE 239	GE 214	GE 208	GE 231

percentile scores. Equipercntile ranks are used to link raw scores from different forms to a common underlying scale in an effort to develop a scale of scores that can be used with all test forms. In this way, the specific form used to test a child becomes arbitrary because differences between forms have been removed through the equating process. In contrast, norm-referenced percentile scores allow one to draw inferences about the performance of a given student relative to the performance of a target reference group. These norm-referenced percentile scores should be applied to the equated scores, and it is, of course, possible to develop multiple sets of reference norms. For example, a student's score could be expressed relative to other students in the same grade, or relative to other students of the same gender in the same grade, or students from a particular grade span. These are different reference groups and the percentile rank for the same equated score would potentially differ across these reference groups. Norm-referenced percentile rankings indicate where students in a given group rank relative to their peers. The same scaled score can map to different normative percentile scores in different reference groups that apply to the same examinee. To understand why this distinction is important, it is instructive to keep in mind that in order to develop normative percentile values one must have a representative sample of the population to which the norms are intended to apply. The same is not true for developing equating tables using equipercntile equating. In fact, with equipercntile equating, what is most needed is information about how the distribution of scores on one measure relates to the distribution of scores for another. It is not uncommon to oversample the tails of the distribution in equating studies to ensure that one has good information about the distribution of scores in the extremes. If one used the same sampling strategy to obtain normative percentile values, one would have to down-weight these extreme scores to offset their having been over-sampled.

## Summary of Linear and Equipercentile Equating

Both linear and equipercentile equating can be used to place passage WCPM scores on a common metric that allows for quick and easy use with all CBM ORF probes from a given assessment system. Linear equating can adjust passages such that they all have a common mean and standard deviation, but as a method it makes the same adjustment at all levels of WCPM. Thus, linear equating assumes that observed WCPM scores have the same relation to true scores at all levels of the underlying fluency ability. This assumption is not made in equipercentile equating. Equipercentile equating relies on observed WCPM percentages and as such can make different score adjustments at different levels of WCPM. For example, it may be that in the center of the ability distribution the relation of observed WCPM and underlying fluency is linear (i.e., a one-unit shift in WCPM results in the same shift in underlying fluency), but observed scores in the tails of the distribution require a different score adjustment relative to scores in the center of the distribution. This differential adjustment was observed in the current empirical example. Specifically, in the tails of the distribution a one-unit shift in the equipercentile rank corresponded to a change in 5–10 WCPM, whereas in the center of the distribution a one-unit shift in the equipercentile rank corresponded to an increase in about 0.50 WCPM. This differential relation of observed WCPM scores at different points in the distribution of WCPM is ignored in linear equating.

Finally, sampling and sample size are important considerations when conducting linear or equipercentile equating studies. Linear equating estimates two parameters per passage—the mean and standard deviation, whereas the estimation of parameters in equipercentile equating is substantially greater. Thus compared to linear equating, equipercentile equating requires substantially larger sample sizes that cover the entire range of WCPM scores in order to reliably estimate parameters across the range of possible WCPM scores. As mentioned, it is not uncommon to oversample the extremes of the distribution when designing a study to use equipercentile equating.

## Latent Variable Equating

One limitation of both linear equating and equipercentile equating is that their focus is explicitly on observed scores. In a sense, these are classical psychometric methods that conceptualize a true score as an unobserved average score for a person hypothetically computed across equivalent forms of a test. An alternative approach conceptually, is to consider the observed scores as imperfect indicators of a latent ability, either in the factor analytic sense of a latent ability, or in the strong true score framework of item response models. Latent Variable (LV) methods are an alternative to observed variable equating methods that reflect this conceptualization of assessments as indicators of latent abilities, although, as we will see, these LV methods are not strictly “equating” methods.

LV equating is the process of equating observed ORF forms based on an unobserved underlying latent ORF ability. Unlike linear equating and equipercentile equating, which are methods for transforming the raw score distributions associated with test forms, LV equating is more an approach to understanding the degree of equivalence across test forms than a method of transforming their raw score distributions. The degree of equivalence or non-equivalence determined through LV equating provides information about the type of transformation that is required to equate the raw scores. Thus, although LV equating is not strict about score equating, the findings from such an LV analysis have implications for the equating of raw scores and the development of constant scales. It is also possible to use the factor scores that result from an LV analysis as a type of equated scale score and we discuss the challenge to develop such an equating table in the final section of this chapter. In presenting LV equating, we note that there are two general approaches to equating by using LVs. Specifically, LV equating can be conducted using either linear or nonlinear relations between observed indicators and the LVs (Stoolmiller, Biancarosa, & Fien, 2013; Zopluoglu, 2013). As with any measurement model, in order to study form equivalence using LV models it is generally necessary to measure students on at least three different test forms measuring the same LV. Following an investigation of the equivalence of observed measures, equating tables can be generated from LV equating similar to equating tables generated from either linear or equipercentile equating (e.g., Betts et al., 2009; Francis et al., 2008). Once these tables are established, a score on a single observed passage can be directly compared to any other equated passage. Thus, like with other methods of equating, the data collection requirements for the equating study are not the same as the data collection requirements for assessing students after forms have been equated. The necessity to use at least three forms is a requirement of the LV equating method. It is not necessary to use at least three forms to assess students once the test forms have been equated, although it may still be desirable to use multiple equated probes at each occasion of measurement to improve the accuracy of measuring students' ORF.

After discussing different approaches to LV equating, we discuss the issues involved in using the relation between the observed scores and latent constructs to establish an equating table and provide an example of such. First, we discuss LV equating and the different approaches to LV equating that have been used with CBM assessments. We discuss both linear and nonlinear LV equating methods and present an empirical example that shows how the measurement models investigated through LV equating methods have implications for the equivalence of observed scores and how this information can subsequently be used to generate equating tables.

## **Linear Latent Variable Equating**

Linear LV equating makes use of constrained factor models in confirmatory factor analysis, or LV structural equation modeling, to extend the traditional linear equating of observed scores. Like traditional linear equating, linear LV

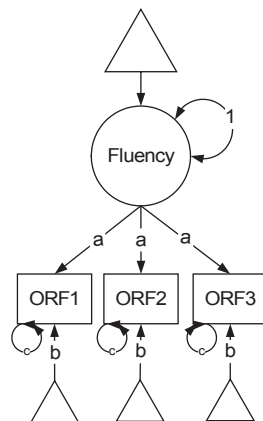


equating can account for and correct differences in means and standard deviations, but does not correct for distributional differences (Cummings, Park, & Bauer Schaper, 2013). However, linear LV equating extends traditional linear equating by assessing the relation between observed scores and latent abilities and using this relation to develop a common scale on which to express scores from individual test forms. Just as in confirmatory factor analysis, in LV equating, a latent factor is assumed to be the cause of the observed scores and, because multiple indicators are used to identify the latent factor, its mean is more precise than the mean of any single indicator. Observed score means can be computed from the factor mean and factor loading in this fashion,  $M = \lambda\alpha + v$ , where  $M$  is the observed mean,  $\lambda$  is the factor loading,  $\alpha$  is the mean of the latent factor, and  $v$  is the measurement intercept.

The process of linear LV equating involves fitting observed ORF scores to a measurement model. Observed ORF indicators are treated as continuous, linear indicators of an underlying fluency ability. Within the LV equating framework, three types of model constraints can be imposed. These constraints imply increasing degrees of equivalence between the different CBM ORF forms. The three types of constraints and the four resulting models are analogous to the models for establishing measurement invariance (e.g., Lubke, Dolan, Kenderman, & Mellenbergh, 2003; Vandenberg & Lance, 2000). The assumptions that these model constraints make and the practical meaning of these constraints will be described below as they relate to the problem of test equating.

Under the most restrictive model, forms are assumed to be parallel. In this model, depicted in Fig. 9.2, factor loadings, measurement intercepts, and residual variances are all constrained to be equal across forms. As shown in the figure, all three observed ORF scores have the same residual variance,  $c$ , the same mean,  $b$ , and are on the same scale,  $a$ , that is, their relation to the underlying construct is identical. Under these model constraints, observed scores on one form are not only equal at the mean, but observed deviations about the mean are comparable between forms. Raw scores on parallel forms are substitutable for one another. That is to say, a person with a given ability has the same expected raw score on any two parallel

**Fig. 9.2** Fully constrained LV equating model. The fully constrained model is depicted here. Parameters that share a common letter are constrained to be equal—a single factor loading, measurement intercept, and residual variance are estimated



forms. Thus, when raw scores are used to make decisions about examinees (e.g., the examinee is “at-risk”), parallel forms make it arbitrary which form is used with a particular examinee. Only with parallel forms is the same raw score expected for the examinee regardless of the form used to assess them, and thus, it is inconsequential to the examinee which form was used to assess them. LV equating allows the test developer to test whether or not different test forms are parallel, and thus, whether or not raw scores are substitutable for one another without loss of information or impact to the examinee.

A less restrictive model would relax the constraint,  $c$ , on residual variances (see Fig. 9.2). This model still assumes that the true scores from the ORF passages are all on the same scale,  $a$ , and have the same mean,  $b$ , but the magnitude of the residual variances is allowed to differ between the ORF forms, which implies that the observed scores are on different scales (i.e., have different variances). In other words, students scoring at a particular value of the LV will have the same *expected* observed scores on different forms. However, because the error variance is different, the distribution of observed scores for these students will be different for different ORF forms, even though those distributions have the same average value.

One way to conceptualize this non-equivalence is to imagine having two regression lines with the exact same slope, but different spreads in the points around the regression lines. The lines depict the relation between the observed score on the vertical axis and the true score (i.e., the score on the factor) on the horizontal axis. Points on the lines depict the expected observed score of all individuals with a given factor score, whereas the points around the line depict how the actual observed scores deviate from the expected scores because of measurement error. The situation being described here is one of coincident lines where the spread of the data points around the two lines is not the same. The coincident lines imply that the expected (i.e., average) observed score is the same on the different forms for people with the same factor score, but the spread of the observed scores around those expectations are different across the different forms.

A still lesser constrained model would remove both the constraint,  $c$ , on residual variances and the constraint,  $b$ , on the observed ORF scores’ measurement intercepts, but the factor loadings,  $a$ , are still held equal. In this model, the ORF true scores are still on the same scale but the true score means and observed score means are allowed to differ across forms. Thinking back to the analogy of two regression lines, we now have two parallel lines, but not two coincident lines. Thus, true scores on one form are simply shifted up or down a constant amount relative to true scores on the other form. The observed score distributions are similarly shifted. ORF forms in this model are considered *essentially tau equivalent*. In other words, because the true scores are the same except for the constant shift, any relation with other measures, such as a measure of reading comprehension, will be the same for the essentially tau equivalent forms, except for the intercept. Practically speaking, the expected observed score for people with the same score on the factor would differ between people taking the two forms, but this difference would be the same throughout the distribution of ability. In that sense, the true scores are not literally equivalent, but they are essentially equivalent.

In a final, fully unconstrained model, all model parameters are allowed to differ across forms. In this case, observed ORF means differ, the spread of scores about those means differ, and the observed scores' relation to the underlying fluency factor differ. Thinking back to the regression analogy, we now have lines that differ in their slopes as well as their intercepts. Thus, the expected observed score differs across forms for people with the same true score. Moreover, the magnitude of this difference in expected scores varies across the distribution of ability. Finally, not only do the expected scores differ between test forms for people with the same true ability, the distributions of observed scores around these expected values also differ across the forms. The one assumption that this model continues to make is that the three forms measure the same unidimensional fluency construct, an assumption that would be evaluated through the examination of fit indices like in a traditional CFA.

### ***Nonlinear Latent Variable Equating Using the Continuous Response Model***

Although the continuous response model (CRM) was proposed as a limiting form of the graded response model over 40 years ago (Samejima, 1973), there have been few applications of the model (Bejar, 1977; Ferrando, 2002; Wang & Zeng, 1998) and only one instance we were able to identify where CRM was applied to fluency data (Zopluoglu, 2013). The CRM can be fit using either limited-information or full-information estimations methods. Because full-information estimation methods are currently unavailable for nonlinear models in standard LV modeling software packages (Zopluoglu, 2013), we will restrict our discussion and empirical examples of the CRM to limited-information estimations methods.

Nonlinear LV equating under the limited-information CRM framework is quite similar to linear LV equating with the primary difference being the handling of the observed data. Under CRM, an observed score  $x$  is considered to be a graded response between 0 and  $m$ , where  $m$  is the maximum possible observed score, and  $y$  is a rescaled observed score such that  $y = \frac{x}{m}$  (Samejima, 1973). Thus, in the context of ORF assessment, the observed scores are rescaled such that an individual's score is divided by the maximum possible score on a given ORF form, resulting in scores between 0 and 1. These rescaled scores are then transformed to logits  $\left( \log \text{it}(y) = \log \left( \frac{y}{1-y} \right) \right)$ , and the covariance matrix of these transformed logits is used as the input data for model fitting (Bejar, 1977; Ferrando, 2002). Once the data are input in this fashion, the same types of linear measurement models described above are used to model the data (Bejar, 1977; Ferrando, 2002). For the sake of comparison, in presenting our empirical example, we will estimate the same four sets of model constraints described above with both linear LV equating and nonlinear CRM LV equating.

## *Latent Variable Equating: Empirical Example*

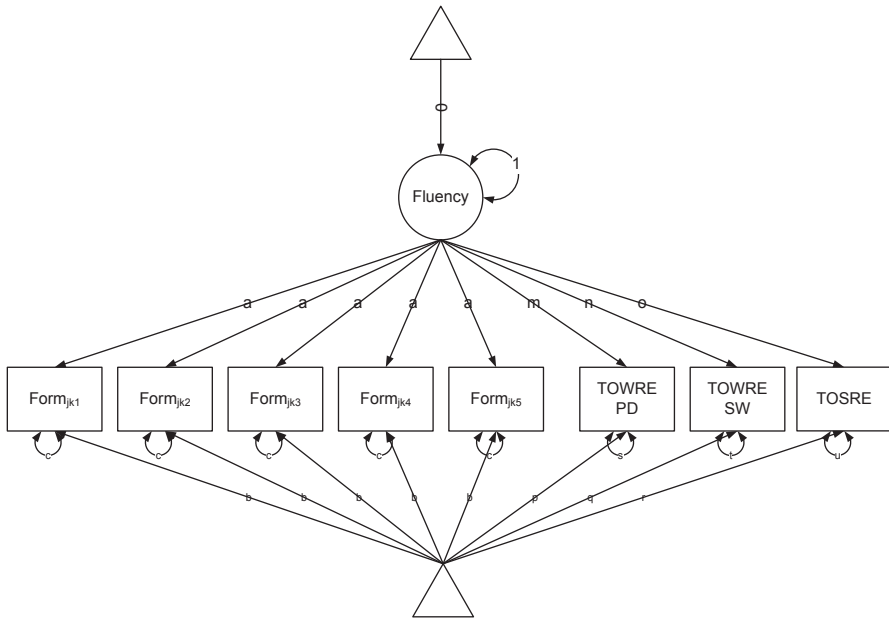
In this section, we present the results of our empirical example. We will begin with presenting the linear LV equating from the most restrictive to least restrictive models. Following this section, the models for the continuous response nonlinear model will be presented.

### **Linear Latent Variable Equating**

In this section, we present the four models described above, moving from the most constrained to the least constrained models. Recall for the design described above that the 35 ORF passages were administered to 6th, 7th, and 8th-grade students where seven passages were semi-spiraled within groups such that the A groups were administered passages 1–7 with 6th-grade students receiving passages 1–5, 7th-grade students receiving passages 2–6, and 8th-grade students receiving passages 3–7. Then the B groups received passages 8–14 and these passages were spiraled in the same fashion. Because common ORF passages are repeated within groups but not between groups, we used three additional assessments to anchor all passages to the same metric. These three assessments were the TOWRE Phonemic Decoding, TOWRE Sight Word, and the TOSRE. All students across all groups and grades received the same version of each of these three assessments. Figure 9.3 presents the measurement model for a single group. Note that the figure portrays the fully constrained model as in Fig. 9.2, but in this figure the factor loadings, measurement intercepts, and residual variances for the TOWRE Phonemic Decoding, TOWRE Sight Word Decoding, and the TOSRE are allowed to differ from the ORF passages. These parameters are allowed to differ because even though these three assessments measure fluency they are not expected to function the same as the ORF passages.

In the first, most constrained model, we fix all factor loadings, measurement intercepts, and residual variances to be the same for all passages. The results for 7th grade, group A are presented in Tables 9.6 and 9.7. Table 9.6 presents the unstandardized parameter estimates for the five ORF passages read by 7th-grade group A, as well as the group's factor mean and variance. As can be seen in Table 9.6, all five reading passages have the same parameter estimates for factor loadings, measurement intercepts, and residual variances. In other words, this model assumes that observed passages are completely parallel—the passages are on the same scale, have the same mean, and same variance. These assumptions can be seen borne out in the relations between the factor scores and observed scores in Table 9.7, as shown below.

Table 9.7 presents expected observed scores on the different forms for individuals who differ in their ability as measured by their having different factor scores. These differences are expressed in the table as being measured in standard deviation units around the factor mean. This section consists of both a description of how these tables of expected observed scores are generated and a presentation of



**Fig. 9.3** LV equating empirical example model. This figure depicts the fully constrained measurement model for the LV empirical example. Factor loadings, measurement intercepts, and residual variances are constrained equal for the fluency forms, but vary for the TOWRE and TOSRE. TOWRE and TOSRE subtests are included to anchor the fluency factor across groups and grades.

the equating results from the LV fluency models. Recall from the equation above, the expected mean for a given passage is  $M = \lambda\alpha + v$ . Thus, to compute the expected mean for a passage (*Koalas* as an example) the factor loading 31.62 is multiplied by the factor mean 0.25 and the measurement intercept 114.29 is added to this product resulting in an expected mean of 122.26 words correct per minute (WCPM). Note that because of constraints on the factor loadings and intercepts, the factor mean and variance need not be constrained to 0 and 1 for scaling. In order to obtain the expected observed score at different values of the fluency factor all that is necessary is to change the value of the fluency factor by adding the desired difference to the factor mean. For example, if one wanted to compute the expected observed score on *Koalas* for individuals with a factor score +2SD from the factor mean, the factor loading 31.62 would be multiplied by this factor score (i.e., the factor mean plus 2 SD, or  $0.25 + 2.13$ ), and this product ( $31.62 \times 2.38$ ) would then be added to the measurement intercept of 114.29 resulting in 189.51 WCPM. Note that the measurement intercept is just the expected observed score for individuals with a score of 0 on the factor. This type of conversion can be done at all points of the fluency factor. With regard to the results for the fully constrained model, expected WCPM scores are the same for all passages at all points of the fluency ability range. Thus, a student with a fluency ability 2SD above the mean would be expected to read 189.51 WCPM on each of the five passages.

**Table 9.6** Grade 7 group A oral reading fluency (ORF) passage parameter estimates: linear LV equating

Parameter	Stories	Model constraint			
		Factor loadings; measurement intercepts; residual variances	Factor loadings; measurement intercepts	Factor loadings	Unconstrained
Factor loadings	Koalas	31.62	32.47	32.01	35.70
	Suni	31.62	32.47	32.01	30.66
	An Unusual Job	31.62	32.47	32.01	34.14
	A Wonderful Friendship	31.62	32.47	32.01	29.44
	The King's Gold	31.62	32.47	32.01	33.43
Measurement intercepts	Koalas	114.29	114.06	119.97	128.10
	Suni	114.29	114.06	122.46	130.60
	An Unusual Job	114.29	114.06	111.34	119.47
	A Wonderful Friendship	114.29	114.06	113.82	121.96
	The King's Gold	114.29	114.06	113.04	121.18
Residual variances	Koalas	297.35	205.45	186.41	164.86
	Suni	297.35	134.09	68.72	68.28
	An Unusual Job	297.35	159.33	133.20	123.88
	A Wonderful Friendship	297.35	140.76	145.63	141.44
	The King's Gold	297.35	158.60	149.94	146.66
Factor mean		0.25	0.27	0.24	0.00
Factor variance		1.13	1.20	1.18	1.00

We have shown how the tabled values of expected observed scores and LV scores can be generated from the results of the factor model. In practice, of course, the desire is to obtain a score on the common scale from the WCPM score on a particular CBM ORF probe. In the fully constrained model, there is a one-to-one mapping of observed scores to scores on the LV. Regardless of which CBM ORF probe was administered, the expected observed score can be found in the table and translated into a score on the LV distribution. As the relation between the LV and the observed score becomes less constrained across different forms, the table begins to appear somewhat more complex.

**Table 9.7** Grade 7 group A: Relation between expected observed scores and factors scores for different linear latent variable (LV) equating models

Constrained factor loadings, measurement intercepts, and residual variances					
Factor score	Koalas	Suni	An Unusual Job	A Wonderful Friendship	The King's Gold
2.38	189.51	189.51	189.51	189.51	189.51
1.85	172.70	172.70	172.70	172.70	172.70
1.32	155.89	155.89	155.89	155.89	155.89
0.78	139.07	139.07	139.07	139.07	139.07
0.25	122.26	122.26	122.26	122.26	122.26
-0.28	105.45	105.45	105.45	105.45	105.45
-0.81	88.63	88.63	88.63	88.63	88.63
-1.34	71.82	71.82	71.82	71.82	71.82
-1.87	55.01	55.01	55.01	55.01	55.01
<i>Constrained factor loadings and measurement intercepts</i>					
2.46	193.84	193.84	193.84	193.84	193.84
1.91	176.09	176.09	176.09	176.09	176.09
1.36	158.34	158.34	158.34	158.34	158.34
0.82	140.58	140.58	140.58	140.58	140.58
0.27	122.83	122.83	122.83	122.83	122.83
-0.28	105.07	105.07	105.07	105.07	105.07
-0.82	87.32	87.32	87.32	87.32	87.32
-1.37	69.56	69.56	69.56	69.56	69.56
-1.92	51.81	51.81	51.81	51.81	51.81
<i>Constrained factor loadings</i>					
2.42	197.42	199.91	188.79	191.26	202.05
1.88	180.01	182.50	171.38	173.86	183.87
1.33	162.60	165.09	153.97	156.45	165.69
0.79	145.19	147.68	136.56	139.04	147.51
0.24	127.78	130.27	119.15	121.63	129.33
-0.30	110.37	112.86	101.74	104.22	111.15
-0.84	92.96	95.45	84.33	86.81	92.98
-1.39	75.56	78.04	66.92	69.40	74.80
-1.93	58.15	60.63	49.52	51.99	56.62
<i>Unconstrained model</i>					
2.00	199.50	191.91	187.76	180.84	188.03
1.50	181.65	176.58	170.69	166.12	171.32
1.00	163.80	161.26	153.62	151.40	154.60
0.50	145.95	145.93	136.54	136.68	137.89
0.00	128.10	130.60	119.47	121.96	121.18
-0.50	110.25	115.27	102.40	107.24	104.46
-1.00	92.40	99.94	85.33	92.52	87.75
-1.50	74.55	84.61	68.26	77.80	71.04
-2.00	56.71	69.28	51.19	63.07	54.32



In the second model, the constraint on residual variances is relaxed. The results of this model can also be seen in Tables 9.6 and 9.7. As seen in Table 9.6, parameter estimates for factor loadings and measurement intercepts are the same across ORF passages but the parameter estimates for residual variances differ. This model assumes that all passages are on the same scale and that the expected observed scores for different passages are identical. However, residual variation about the mean is allowed to differ across passages. Note that this differential residual variation has no bearing on the expected observed score on different forms at any given value of the factor score, for example, on any form the expected observed score is the same for people scoring 2SD below the factor mean, just as it is the same across forms for people scoring at the factor mean, or for people scoring at 2SD above the factor mean. Expected observed scores differ for people with different scores on the factor, but expected observed scores are the same on each form for any given set of individuals with the same score on the factor. As shown in Table 9.7, individuals with a factor score of 1.91 have an expected score of 176.09 WCPM on any given form, whereas individuals with a score of 0.82 on the factor would be expected to score 140.58 on any given form. These expectations would hold for anyone with a score of 1.91 or 0.82 on the factor, respectively. Because the residual errors are assumed to have an expected value of 0, relaxing the constraint of equal residual variances has no impact on the expected observed scores for different passages at different points on the fluency factor.

This lack of impact on the expected observed scores is shown in Table 9.7. Like the results of the fully constrained model, in this less constrained model all passages have the same expected value at any given point on the fluency factor. Although expected observed scores differ by an amount determined by the factor loading as one moves away from the factor mean, this difference is the same for each form and thus no difference in expected scores is found across forms. Thus, if the assumptions of this model are reasonable, then passages are interchangeable to a certain extent. In so far as a student's expected score is the same across forms, it is somewhat arbitrary which form a student is administered. However, because the error associated with the student's score is not the same across forms, the possibility of having an observed score that differs greatly from the student's expected score is not the same for the two forms. Generally speaking, it is most fair if all students are tested on forms that are equally precise in estimating their true ability. That is, it may not be enough to know that expected scores are equal across the forms if one form is associated with substantially more error than another form.

In the third model, the constraint on measurement intercepts is relaxed. As shown in Table 9.6, factor loadings of the five passages are identical but the measurement intercepts and residual variances differ. This model assumes that true scores for passages are on the same scale (factor loadings constrained equal), but allows them to have different mean values, and the residual variances associated with observed scores are allowed to differ. The different estimates for intercepts affects both the mean of the true scores and the mean of the observed scores, whereas the differences in residual variances affect the variances of the observed scores, but not the variances of the true scores. In other words, this model assumes that passages are essentially

tau equivalent—true scores differ across forms by a constant amount at all points of the latent ability continuum. This equivalence can be seen in Table 9.7. As shown in the table, the expected values of the ORF passages differ for all passages at all points along the range of fluency. However, the magnitude of the difference between any two passages is the same at all points of the fluency ability range. For example, if one compares the difference in expected observed scores for *Koalas* and *Suni* in Table 9.7 for the model with only constrained factor loadings, one finds that the difference in expected observed scores is always 2.49 (e.g.,  $199.91 - 197.42 = 2.49$  for individuals with a factor score of 2.42,  $182.50 - 180.01 = 2.49$  for individuals with a factor score of 1.88, etc.). This common magnitude in the difference between the forms occurs because of the computation of the expected mean  $M = \lambda\alpha + v$ . For all forms  $\lambda\alpha = 32.01 \times .24$  and forms only differ in their measurement intercept,  $v$ . For *Koalas* and *Suni*, this difference in intercepts is 2.49 (i.e.,  $122.46 - 119.97 = 2.49$ ), as shown in Table 9.6 for the model with constrained factor loadings. The difference in intercepts coupled with the constraint of equal slopes translates into a constant difference in the expected observed scores (i.e., the true scores) for any two forms across the range of the latent factor.

In the final, unconstrained model, there are no constraints placed on any ORF passage parameters. However, for model identification purposes the factor mean and variance are fixed to 0 and 1, respectively. This model allows all passages to be on different scales have different measurement intercepts and different residual variances. As shown in Table 9.6, all model parameters differ across the ORF passages. These differences in the scales and intercepts are illustrated in the tables of expected observed scores in Table 9.7. For example, *Koalas* and *Suni* differ by about 2.5 WCPM at the mean of the passages, but they differ by about 7.5 WCPM at 2SD above the mean.

In summary, LV equating can be used to test different sets of model constraints that translate into different degrees of equivalence across test forms.<sup>1</sup> In the two models, where factor loadings and measurement intercepts were constrained to be equal, the model results and tables of expected observed scores illustrate how ORF forms are either parallel, when all parameters are equal across forms, or produce interchangeable expected observed scores when only slopes and intercepts are equal across forms. Of course, because of differences in precision, forms are not fully interchangeable in the latter case. In the model where only factor loadings are constrained equal, the model allows the means of the expected observed scores for the ORF passages to differ, but holds the scale for the expected observed scores of passages constant. Although this constraint ensures that regression relations with other variables will be the same for different passages, the constant difference in expected observed scores could result in different decisions when scores are compared to benchmarks if the shift in the scale means is ignored. In the final, unconstrained model, scales and measurement intercepts differ across forms, along with residual

<sup>1</sup> The appropriate model for the data depends on tests of invariance, which are typically carried out by examining the  $\chi^2$  test of model fit, comparing the  $-2$  log likelihood of a more constrained nested model with that of a less constrained model.

variances, which is clearly seen in the differences in the tables of expected observed scores. Finally, it is important to note that these are all nested models and as such the appropriateness of the model constraints is directly testable by comparing the likelihood ratios of the models. Researchers would typically perform these tests when investigating differences between CBM ORF probes. The purpose here was to demonstrate LV equating, not to explore the features of these particular CBM ORF probes, the examination and discussion of which are beyond the scope of this section. There are many excellent references on the comparison and testing of nested models in LV modeling contexts (e.g., Mueller & Hancock 2008).

### Nonlinear Latent Variable Equating Using the Continuous Response Model

Like in the linear LV equating example above, this section contains the four CRM models beginning with the most constrained model and ending with the least constrained model. Recall that once observed scores are put on a metric from 0 to 1 and then converted to logits, the means and covariances can be modeled just as they would for the linear LV equating models. As such, the model portrayed in Fig. 9.3 is still the example for a measurement model for a single group and this model continues to use the TOWRE Phonemic Decoding, TOWRE Sight Word, and the TOSRE as anchor assessments across groups and grades.

In the first CRM model, all factor loadings, measurement intercepts, and residual variances are fixed to be the same for all passages. The results for 7th grade, group 1 continue to be used as an example and these results shown in Tables 9.8 and 9.9. Again, Table 9.8 presents the unstandardized parameter estimates for the five ORF passages and the group's factor mean and variance. As shown in Table 9.8, all five reading passages have the same parameter estimates for factor loadings, measurement intercepts, and residual variances. However, because these data were transformed in a non-linear way prior to modeling, these constraints do not result in equivalent expected observed scores once the data are scaled back into their original metric. This lack of parallelism from the fully constrained model can be seen in the table of expected observed scores, which are presented in Table 9.9.

Table 9.9 presents the expected observed scores for the different CRM models for these five passages for a range of fluency scores  $\pm 2SD$  to  $+2SD$  from the factor mean. Expected observed scores are presented in both the logit metric and the WCPM metric. The construction of these tables is somewhat different than the tables constructed in Table 9.7 for the linear LV equating models. The first step in creating the tables is the same as above—computing the expected observed score for a given passage with the equation  $M = \lambda\alpha + v$ . Using the *Koalas* passage as an example, the factor loading 0.69 is multiplied by the factor mean 0.42 and the measurement intercept  $-0.20$  is added to this product resulting in an expected observed score mean of 0.09. However, these model parameters and the resulting expected means are on a logit scale and not in WCPM. The expected means must then be converted back into a WCPM metric. To be put on a WCPM metric, the expected score must go through a three-step process. In the first step, the exponent of the expected score  $M$  is taken,  $e^M$ . In the current example, the exponent of 0.09 is 1.10. Next,

**Table 9.8** Grade 7 group A: Oral reading fluency (ORF) passage parameter estimates: nonlinear continuous response model

Parameter	Stories	Model constraint			
		Factor loadings; measurement intercepts; residual variances	Factor loadings; measurement intercepts	Factor loadings	Unconstrained
Factor loadings	Koalas	0.69	0.68	0.68	0.81
	Suni	0.69	0.68	0.68	0.70
	An Unusual Job	0.69	0.68	0.68	0.79
	A Wonderful Friendship	0.69	0.68	0.68	0.73
	The King’s Gold	0.69	0.68	0.68	0.77
Measurement intercepts	Koalas	-0.20	-0.27	-0.39	0.00
	Suni	-0.20	-0.27	-0.35	0.04
	An Unusual Job	-0.20	-0.27	-0.45	-0.06
	A Wonderful Friendship	-0.20	-0.27	-0.35	0.04
	The King’s Gold	-0.20	-0.27	-0.41	-0.02
Residual variances	Koalas	0.18	0.09	0.09	0.08
	Suni	0.18	0.03	0.03	0.03
	An Unusual Job	0.18	0.05	0.05	0.04
	A Wonderful Friendship	0.18	0.05	0.05	0.05
	The King’s Gold	0.18	0.05	0.05	0.05
Factor mean		0.42	0.42	0.63	0.00
Factor variance		1.16	1.25	1.23	1.00

this resulting value  $y$  must be converted to a probability with the equation  $\frac{p}{1-p} = y$ , or  $p = \frac{y}{1+y}$ . In this example,  $\frac{1.10}{1+1.10} = 0.52$ . Finally, in the third step, the probability is put back into the WCPM metric with the equation  $p = \frac{x}{\max}$ , or  $x = p * \max$ , where  $\max$  equals the passage maximum used to create the original probabilities. In this example, the passage maximum is 217 WCPM. Thus, a probability of 0.52 results in a WCPM score of 113.5. This process can be repeated across the range

**Table 9.9** Grade 7 group A: Relation between expected observed scores and factor scores for different nonlinear continuous response models

<i>Constrained factor loadings, measurement intercepts, and residual variances: logits</i>					
Factor score	Koalas	Suni	An Unusual Job	A Wonderful Friendship	The King's Gold
2.57	1.58	1.58	1.58	1.58	1.58
2.03	1.21	1.21	1.21	1.21	1.21
1.49	0.84	0.84	0.84	0.84	0.84
0.95	0.46	0.46	0.46	0.46	0.46
0.42	0.09	0.09	0.09	0.09	0.09
-0.12	-0.28	-0.28	-0.28	-0.28	-0.28
-0.66	-0.65	-0.65	-0.65	-0.65	-0.65
-1.20	-1.02	-1.02	-1.02	-1.02	-1.02
-1.74	-1.39	-1.39	-1.39	-1.39	-1.39
<i>Constrained factor loadings and measurement intercepts: logits</i>					
2.65	1.52	1.52	1.52	1.52	1.52
2.09	1.14	1.14	1.14	1.14	1.14
1.53	0.77	0.77	0.77	0.77	0.77
0.97	0.39	0.39	0.39	0.39	0.39
0.42	0.01	0.01	0.01	0.01	0.01
-0.14	-0.36	-0.36	-0.36	-0.36	-0.36
-0.70	-0.74	-0.74	-0.74	-0.74	-0.74
-1.26	-1.12	-1.12	-1.12	-1.12	-1.12
-1.82	-1.49	-1.49	-1.49	-1.49	-1.49
<i>Constrained factor loadings: logits</i>					
2.84	1.54	1.57	1.47	1.57	1.51
2.29	1.16	1.20	1.10	1.20	1.14
1.74	0.79	0.82	0.72	0.83	0.76
1.18	0.41	0.45	0.35	0.45	0.39
0.63	0.04	0.08	-0.02	0.08	0.01
0.08	-0.33	-0.30	-0.40	-0.30	-0.36
-0.48	-0.71	-0.67	-0.77	-0.67	-0.73
-1.03	-1.08	-1.04	-1.14	-1.04	-1.11
-1.58	-1.45	-1.42	-1.52	-1.42	-1.48
<i>Unconstrained model: logits</i>					
2.00	1.62	1.44	1.52	1.51	1.51
1.50	1.22	1.09	1.12	1.14	1.13
1.00	0.81	0.74	0.73	0.78	0.74
0.50	0.41	0.39	0.34	0.41	0.36
0.00	0.00	0.04	-0.06	0.04	-0.02
-0.50	-0.40	-0.31	-0.45	-0.33	-0.41
-1.00	-0.81	-0.66	-0.85	-0.69	-0.79
-1.50	-1.21	-1.01	-1.24	-1.06	-1.17
-2.00	-1.61	-1.36	-1.64	-1.43	-1.55
<i>Constrained factor loadings, measurement intercepts, and residual variances: WCPM</i>					
2.57	179.93	181.58	172.46	167.49	172.46
2.03	167.07	168.61	160.14	155.52	160.14

**Table 9.9** (continued)

Constrained factor loadings, measurement intercepts, and residual variances: logits					
Factor score	Koalas	Suni	An Unusual Job	A Wonderful Friendship	The King's Gold
1.49	151.39	152.78	145.11	140.92	145.11
0.95	133.24	134.47	127.72	124.03	127.72
0.42	113.51	114.56	108.80	105.67	108.80
-0.12	93.44	94.30	89.57	86.98	89.57
-0.66	74.37	75.06	71.29	69.23	71.29
-1.20	57.39	57.92	55.01	53.42	55.01
-1.74	43.11	43.51	41.32	40.13	41.32
<i>Constrained factor loadings and measurement intercepts: WCPM</i>					
2.65	178.04	179.68	170.66	165.73	170.66
2.09	164.53	166.05	157.71	153.16	157.71
1.53	148.15	149.52	142.01	137.91	142.01
0.97	129.38	130.57	124.01	120.43	124.01
0.42	109.21	110.22	104.68	101.66	104.68
-0.14	89.00	89.82	85.31	82.84	85.31
-0.70	70.09	70.74	67.18	65.25	67.18
-1.26	53.52	54.01	51.30	49.82	51.30
-1.82	39.81	40.17	38.16	37.06	38.16
<i>Constrained factor loadings: WCPM</i>					
2.84	178.54	181.32	169.18	167.30	170.34
2.29	165.26	168.20	155.99	155.22	157.42
1.74	149.15	152.21	140.11	140.48	141.80
1.18	130.64	133.73	122.06	123.44	123.92
0.63	110.68	113.67	102.82	104.95	104.74
0.08	90.58	93.33	83.65	86.19	85.51
-0.48	71.66	74.07	65.82	68.42	67.50
-1.03	54.98	56.99	50.26	52.64	51.68
-1.58	41.08	42.68	37.40	39.43	38.55
<i>Unconstrained model: WCPM</i>					
2.00	181.21	177.19	170.64	165.45	170.36
1.50	167.44	164.02	157.00	153.16	157.10
1.00	150.33	148.36	140.36	138.33	141.00
0.50	130.36	130.64	121.28	121.37	122.57
0.00	108.72	111.69	100.93	103.12	102.86
-0.50	87.06	92.61	80.81	84.73	83.22
-1.00	67.04	74.52	62.37	67.38	65.02
-1.50	49.86	58.34	46.59	52.02	49.23
-2.00	36.03	44.59	33.88	39.13	36.30

of the fluency factor by first computing expected scores on the logit scale, just as with linear LV equating, and then applying the three-step conversion process to re-express those values on the original scales. With regard to the results for the fully constrained CRM model, WCPM are the same only if the passage maximum is the

same, as was the case for *An Unusual Job* and *The King's Gold*, but as the passage maximums differed, so did all expected WCPM across the range of the fluency factor. Also shown in the table, the scales of passages that do not have a common maximum differ across passages. The key point in thinking about these constrained models in the case of nonlinear LV modeling is to keep in mind that the equivalence imposed by the model applies to the transformed metric, not the original WCPM metric. Generally speaking, equivalency constraints imposed in LV modeling are not scale invariant. That is, they are tests of equivalence in the metric in which the constraint is applied and their applicability in one metric does not imply that they will hold in a different metric. This caveat is not unique to nonlinear LV models.

In the next model, the constraint on residual variances is removed. The parameter estimates and expected observed scores for this model can also be shown in Tables 9.8 and 9.9. Like the results of the previous model, in this less constrained model the expected scores on the logit scale are identical at all points across the range of the fluency factor, but only ORF passages with common passage maximums will have WCPM scores that are equal and on the same scale. Thus, the only passages that are interchangeable under these model assumptions are those with common maximums. All other passages would need to be equated to one another or back to a single common metric. Again, the caution raised above regarding unequal precisions must be kept in mind for models with equal true scores, but unequal residual variances.

In the next model, the constraint on measurement intercepts is relaxed. As shown in Table 9.8, factor loadings are the only parameter that is identical across the five passages. On the logit scale, this model assumes that passages are essentially tau equivalent—they are on the same scale but have different intercepts, but they are not essentially tau equivalent on the WCPM scale. This can be clearly shown in Table 9.9 where ORF passages' WCPM differ in both the expected observed scores and the scale.

In the unconstrained model, there are no constraints on any parameters. In this model, passage scales, measurement intercepts, and residual variances differ on the logit scale, as well as the WCPM scale. The results for this model shown in Tables 9.8 and 9.9.

### ***Conclusions Regarding the Use of Nonlinear Latent Variable Models in Equating***

In conclusion, CRM equating can be performed on the same types of models as linear LV equating. Model parameters and the resulting expected observe scores for the passages follow the same patterns as those in the linear LV equating models, but the implied equivalence on “expected observed scores” applies to the transformed metric and does not hold once the logits are transformed back to WCPM. Under the CRM framework, even the most restrictive models only indicate that passages are interchangeable when the passage maximums are the same. Otherwise, the resulting



equating tables will have WCPM that differ in both expected observed scores and scales. Further, in models where measurement intercepts, as well as factor loadings, are not constrained, then passages with common maximums cease to be identical across the fluency ability range. Thus, even though the measurement models are identical between linear LV equating and CRM equating and the results for the CRM on the logit scale follow the same pattern as the results of linear LV equating, the results on a WCPM scale are quite different between the two modeling frameworks.

## Converting Raw Scores to Factor Scores

A major contribution of the linear and non-linear factor models presented above is in allowing one to determine the kind of equating that is optimal for the data (Stoolmiller et al. 2013). As Tables 9.7 and 9.9 show, the different constraints imply different degrees of equivalence across the expected observed scores. When tests are parallel, raw scores are equivalent in that expected observed scores are equal and the precision with which true scores are estimated is equal. When the unconstrained factor model provides the best fit to the data, either equipercentile or linear equating is needed depending on the shape of observed score distributions. It is tempting to view Tables 9.7 and 9.9 as providing the transformation from raw scores to factor scores, i.e., to view these tables as equating tables that allow one to place observed scores onto an equated scale. However, they do not serve that purpose. The reason these tables cannot be used to find the factor score associated with a particular observed test score rests on an important, but somewhat overlooked problem in factor models, namely the problem of obtaining factor scores from observed variable scores. As the reader will no doubt recall, the values in Tables 9.7 and 9.9 were produced by taking factor scores and converting them to expected observed scores. Practically speaking the problem stems from the fact that Tables 9.7 and 9.9 show us the expected observed score for a person with a particular factor score. In test equating, we have the observed score; what we desire is the factor score. What we need then is not a table that takes us from the factor score to the observed score, but from the observed score to the expected factor score. Because the observed scores are measured with error, the expected factor score associated with a particular observed score is not the same as the factor score that produces that observed score as an expected observed score. Put another way, returning to our regression analogy used to explain the different model constraints, in regression, the slope of the line predicting Y from X is not generally the reciprocal of the line predicting X from Y, except when X and Y are measured without error. That is, if  $Y_i = b_0 + b_1 X_i + e_p$ , then why is not also true that  $X_i = b'_0 + b'_1 Y_i + e'_p$ , where  $b'_1 = 1/b_1$  and  $b'_0 = -b_0/b_1$ ? The answer is error in Y. If Y were measured without error, then, indeed, we could simply use Tables 9.7 and 9.9 to obtain expected factor scores from individual instances of expected observed WCPM scores. Unfortunately, the observed scores contain errors, and so we need to determine the expected factor score associated with a given

expected observed score when the distribution of observed scores varies around the expected observed scores due to measurement error. It is possible to make this determination using what are known as the factor score regression coefficients in factor analysis, or factor score estimation in standard psychometric software, such as MPlus. These factor scores will be approximately normally distributed and will have a mean of 0 and standard deviation of 1.0. We can use these factor scores to develop a normally distributed scale score and then develop equating tables by using regression to find the expected scaled score associated with each observed score.

In Table 9.10, we provide such an equating table based on the unconstrained linear factor model. In this case, the scaled score was constructed to have a mean of 128 and a standard deviation of 15. These choices were arbitrary, but were chosen because the mean of the five observed scores for grade 7 students was approximately 128. After a student is tested on one of the five passages, the examiner would simply find the column associated with the passage read by the student, and then find the obtained row associated with the raw score in the column for that form, and convert the raw score to the Scaled Score number that appears in the first column and the same row. The process of generating the table is a bit beyond the scope of this chapter, but suffice it to say that the results will either parallel mean, linear, or equipercentile equating with observed scores depending on the factor model that was used to develop the factor score regression coefficients.

## Conclusions

ORF has a long-standing research base as a way to measure student growth in reading through the use of curriculum-based measurement (CBM) techniques. CBM has provided special education teachers with a reliable and valid measurement tool that can be quickly administered and allows for both teachers and students to chart progress in several academic content areas. With several federal legislative mandates in place, general education has increasingly turned to CBM as an efficient way to monitor student progress. Given the growing popularity of this type of assessment, it is important for researchers to find ways to ensure that the results indicate true growth in reading ability and not artifacts resulting from form effects by using forms of different difficulty at different assessment times, or with different students.

This chapter provides researchers and educators with alternative methods for translating raw scores (simple words correct per minute) into scaled scores with more desirable measurement properties than observed scores, and thus reducing the variability in raw scores that results from differences in form difficulty. The expression of observed WCPM scores on a constant metric provides classroom teachers a more accurate and stable metric on which to quantify changes in reading ability over time. Research in this area is important not only to the development of better assessments, but also to assist teachers and administrators in their use and interpretation of test data, and in their efforts to convey the meaning of test data to all stakeholders including parents, school boards, and legislative bodies.

**Table 9.10** Grade 7 group A ORF scale score table for equating raw scores based on an unconstrained linear factor model

Scale score	Koalas	Suni	An Unusual Job	A Wonderful Friendship	The King's Gold
165		219			
161		217–218	208		
160		214–216	206–207	202	
159		211–213	204–205	201	207–208
158	217	207–210	202–203	199–200	203–206
157	215–216	204–206	200–201	196–198	201–202
156	213–214	201–203	198–199	193–195	199–200
155	211–212	198–200	196–197	190–192	197–198
154	208–210	195–197	193–195	188–189	195–196
153	206–208	192–194	190–192	186–187	192–194
152	203–205	189–191	188–189	184–185	189–191
151	200–202	187–188	185–187	182–183	186–188
150	196–199	185–186	182–184	180–181	183–185
149	193–195	183–184	178–181	177–179	181–182
148	190–192	181–182	176–177	174–176	179–180
147	187–189	179–180	174–175	171–173	177–178
146	183–185	177–178	172–173	169–170	174–176
145	181–182	175–176	169–171	167–168	171–173
144	179–180	173–174	165–168	164–166	168–170
143	176–178	170–172	163–164	162–163	166–167
142	173–175	167–169	162	160–161	164–165
141	170–172	165–166	160–161	158–159	162–163
140	167–169	163–164	158–159	156–157	159–160
139	164–166	161–162	155–157	153–155	156–158
138	162–163	158–160	152–154	151–152	154–155
137	160–161	156–157	149–151	149–150	152–153
136	157–159	154–155	146–148	147–148	149–151
135	154–156	152–153	144–145	145–146	146–148
134	152–153	150–151	141–143	142–144	143–145
133	149–151	148–149	139–140	140–141	141–142
132	146–148	146–147	137–138	137–139	138–140
131	143–145	144–145	134–136	135–136	135–137
130	141–142	141–143	131–133	133–134	132–134
129	138–140	139–140	129–130	130–132	130–131
128	135–137	137–138	126–128	128–129	128–129
127	133–134	134–136	124–125	126–127	126–127
126	130–132	132–133	121–123	124–125	123–125
125	127–129	130–131	119–120	122–123	120–122
124	124–126	128–129	116–118	119–121	118–119
123	122–123	125–127	113–115	117–118	116–117
122	119–121	123–124	111–112	114–116	113–115
121	117–118	121–122	108–110	112–113	110–112

**Table 9.10** (continued)

Scale score	Koalas	Suni	An Unusual Job	A Wonderful Friendship	The King's Gold
120	114–116	119–120	106–107	110–111	108–109
119	111–113	117–118	103–105	107–109	105–107
118	107–110	114–116	101–102	105–106	103–104
117	105–106	112–113	98–100	103–104	101–102
116	103–104	110–111	96–97	100–102	98–100
115	100–102	108–109	93–95	98–99	95–97
114	98–99	106–107	90–92	96–97	93–94
113	95–97	104–105	88–89	93–95	91–92
112	92–94	101–103	85–87	91–92	89–90
111	89–91	99–100	83–84	88–90	86–88
110	86–88	96–98	81–82	86–87	83–85
109	84–85	93–95	78–80	84–85	80–82
108	81–83	91–92	76–77	82–83	78–79
107	77–80	89–90	73–75	79–81	75–77
106	76	87–88	70–72	77–78	72–74
105	73–75	84–86	68–69	75–76	70–71
104	71–72	82–83	65–67	73–74	68–69
103	68–70	80–81	63–64	70–72	65–67
102	64–67	78–79	61–62	68–69	62–64
101	61–63	76–77	59–60	66–67	
100	60	74–75			
99	59	72–73			
98		70–71		59	
97		68–69			51
96		66–67		55	
95		65			
93					41
92	37				
91			34		
90				40	
89			28		
86		45			
83	13				

## References

- Ardoin, S., Suldo, S., Witt, J., Aldrich, S., & McDonald, E. (2005). Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly*, *20*(1), 1–22. doi:10.1521/scpq.20.1.1.64193.
- Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, *21*(3), 285–301. doi:10.1016/S0271-5309(01)00005-2.
- Barth, A., Stuebing, K. K., Fletcher, J. M., Cirino, P. T., Francis, D. J., & Vaughn, S. (2012). Reliability and validity of the median score when assessing the oral reading fluency of middle grade readers. *Reading Psychology*, *33*(1–2), 133–161. doi: 10.1080/02702711.2012.631863.

- Begeny, J. C., & Greene, D. J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools, 51*(2), 198–215. doi:10.1002/pits.21740.
- Bejar, I. I. (1977). An application of the continuous response level model to personality measurement. *Applied Psychological Measurement, 1*, 509–521.
- Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *School Psychology, 47*, 1–17. doi:10.1016/j.jsp.2008.09.001.
- Blommers, P. J., & Forsyth, R. A. (1977). *Elementary statistical methods in psychology and education* (2nd ed.). Boston: Houghton Mifflin.
- Compton, D. L., Appleton, A. C., & Hosp, M. K. (2004). Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average to poor decoders. *Learning Disabilities Research & Practice, 19*(3), 176–184. doi:10.1111/j.1540-5826.2004.00102.x.
- Cummings, K. D., Atkins, T., Allison, R., & Cole, C. (2008). Response to Intervention: Investigating the new role for special educators. *Teaching Exceptional Children, 40*(4), 24–31.
- Cummings, K. D., Park, Y., & Bauer Schaper, H. A. (2013). Form effects on DIBELS Next oral reading fluency progress monitoring passages. *Assessment for Effective Intervention, 38*(2), 91–104. doi:10.1177/1534508412447010.
- Dale, E., & Chall, J. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin, 27*(2), 37–54.
- Deno, S. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184–192. doi:10.1177/00224669030370030801.
- Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*(4), 507–524.
- Ferrando, P. J. (2002). Theoretical and empirical comparisons between two models for continuous item responses. *Multivariate Behavioral Research, 37*, 521–542. doi:10.1207/S15327906MBR3704\_05.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233.
- Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (Eds.). (2007). *Learning disabilities: From identification to intervention*. New York: Guilford.
- Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*(3), 315–342. doi:10.1016/j.jsp.2007.06.003.
- Francis, D. J., Barth, A., Cirino, P., Reed, D. K., & Fletcher, J. M. (2010). *Texas middle school fluency assessment, version 2.0*. Houston: University of Houston/Texas Education Agency.
- Fry, E. (1968). A readability formula that saves time. *Journal of Reading, 11*(7), 513–516.
- Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*(6), 488–501.
- Fuchs, L. S., & Stecker, P. M. (2003). *Scientifically based progress monitoring*. Washington, DC: National Center on Student Progress Monitoring. <http://www.studentprogress.org/library/Presentations/ScientificallyBasedProgressMonitoring.pdf>. Accessed 20 Nov 2014.
- Fuchs, L. S., Fuchs, D. F., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239–256. doi:10.1207/S1532799XSSR0503\_3.
- Good, R. H., & Kaminski, R. A. (2002a). Dynamic indicators of basic early literacy skills (2000–2003). <http://dibels.uoregon.edu/>. Accessed 20 Nov 2014.
- Good, R. H., & Kaminski, R. A. (2002b). *DIBELS oral reading fluency passages for first through third grades (technical report no. 10)*. Eugene: University of Oregon.
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analysis of text characteristics. *Educational Researcher, 40*, 223–234.
- Gunning, R. (1952). *The technique of clear writing*. New York: McGraw-Hill.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage.
- Hammill, D. D., Wiederholt, J. L., & Allen, A. E. (2006). *Test of silent contextual reading fluency*. Austin: Pro-Ed.

- Hiebert, E. (2002). Standards, assessments, and text difficulty. In A. Farstrup & S. Samuels (Eds.), *What research has to say about reading instruction* (pp. 337–369). Newark: International Reading Association.
- Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New York: Academic.
- Kame'enui, E. J., & Simmons, D. C. (2001). Introduction to this special issue: The DNA of reading fluency. *Scientific Studies of Reading*, 5(3), 203–210. doi:10.1207/S1532799XSSR0503\_1.
- Klare, G. R., & Buck, B. (1954). *Your reader: The scientific approach to readability*. New York: Hermitage House.
- Kolen, M. J., & Brennan, R. L. (1995). *Standard errors of equating*. New York: Springer.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.
- Lubke, G. H., Dolan, C. V., Kenderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence*, 31, 543–566. doi:10.1016/S0160-2896(03)00051-5.
- Madelaine, A., & Wheldall, K. (2004). Curriculum-based measurement of reading: Recent advances. *International Journal of Disability, Development and Education*, 51(1), 57–82.
- Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York: Guilford.
- Mathes, P. G., Torgesen, J. K., & Herron, J. *Continuous monitoring of early reading skills (CMERS) (2008) [Computer software]*. San Rafael: Talking Fingers, Inc.
- McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading*, 22(1), 639–646.
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale: Lawrence Erlbaum.
- Mueller, R. O., & Hancock, G. R. (2008). 32 best practices in structural equation modeling. In J. Osborne (Ed.), *Best practice in quantitative methods* (pp. 488–508). NY: Sage.
- National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development.
- Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. New York: Student Achievement Partners.
- Petscher, Y., & Kim, Y. S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology*, 49(1), 107–129. doi:10.1016/j.jsp.2010.09.004.
- Powell-Smith, K. A., & Bradley-Klug, K. L. (2001). Another look at the “C” in CBM: Does it really matter if curriculum-based measurement probes are curriculum-based? *Psychology in the Schools*, 38(4), 299–312. doi:10.1002/pits.1020.
- Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, 38, 203–219.
- Shinn, M. R. (2002). Best practices in using curriculum-based measurement in a problem solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (pp. 671–698). Bethesda: National Association of School Psychologists.
- Shinn, M. R., Rosenfield, S., & Knutson, N. (1989). Curriculum-based assessment: A comparison of models. *School Psychology Review*, 18(3), 299–316.
- Snow, C., Burns, M., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Spache, G. (1953). A new readability formula for primary grade reading materials. *The Elementary School Journal*, 53(7), 410–413.
- Sprick, M., Howard, L. M., & Fidanque, A. (1998). *Read well: Critical foundations in primary reading*. Longmont: Sopris West.
- Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2007). *The Lexile framework for reading technical report*. Durham: MetaMetrics, Inc.

- Sticht, T. G. (1973). Research toward the design, development and evaluation of a job-functional literacy program for the US Army. *Literacy Discussion*, 4(3), 339–369.
- Stoolmiller, M., Biancarosa, G., & Fien, H. (2013). Measurement properties of DIBELS oral reading fluency in grade 2: Implications for equating studies. *Assessment for Effective Intervention*, 39(2), 76–90. doi:10.1177/1534508412456729.
- Swanson, C. E., & Fox, H. G. (1953). Validity of readability formulas. *Journal of Applied Psychology*, 37(2), 114–118. doi:10.1037/h0057810.
- Tekfi, C. (1987). Readability formulas: An overview. *Journal of Documentation*, 43(3), 257–269. doi:10.1108/eb026811.
- Texas Education Agency (TEA), University of Texas, Health Science Center (UTHSC), and University of Houston. (2010). *The Texas Primary Reading Inventory (TPRI)*. Baltimore: Brookes Publishing.
- Torgesen, J. K., Wagner, R., & Raschote, C. (2001). *Test of word reading efficiency*. Austin: Pro-Ed.
- Vandenberg, R. J., & Lance, C. E., (2000). A review and synthesis of measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(4), 4–70. doi:10.1177/109442810031002.
- Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *Test of sentence reading efficiency and comprehension*. Austin: Pro-Ed.
- Wang, T., & Zeng, L. (1998). Item parameter estimation for a continuous response model using an EM algorithm. *Applied Psychological Measurement*, 22, 333–344. doi:10.1177/014662169802200402.
- Zoplouglu, C. (2013). A comparison of two estimation algorithms for Samejima's continuous IRT model. *Behavioral Research Methods*, 45, 54–64. doi:10.3758/s13428-012-0229-6.



## Part III

# Advanced Research Methods

With an assessment selected, psychometrics evaluated, and scores obtained from data collection, a natural question that may emerge is, “What are the new methods for analyzing fluency data?” This part contains three chapters that focus on emerging statistical methods in fluency research, all of which are critical to contemporary research questions. Chapter 10 by McCoach and colleagues provides a discussion of growth curve analysis using structural equation modeling. In this chapter, the authors demonstrate different forms of the individual growth curve and how to model change with multiple outcomes. Chapter 11 by Logan and colleagues presents an introduction to mixture models via latent class analysis. This technique uses a categorical latent variable to test if observed measures identify one or more groups of responders. Chapter 12 by Petscher and colleagues introduces the reader to latent change score analysis, a model which views growth as a function of average change over time as well as a function of auto-regressive and coupling effects. All the three chapters use the same data set to illustrate the technique. The data and code are available at <http://myweb.fsu.edu/ymp5845/>.

# Chapter 10

## Using Individual Growth Curves to Model Reading Fluency

D. Betsy McCoach and Huihui Yu

Longitudinal research generally involves the collection of data from the same people across multiple time points. As longitudinal studies collect repeated “panels” across multiple waves of data collection, researchers often refer to the data collected from such studies as longitudinal data, repeated measures data, or panel data (Frees, 2004; Hsiao, 2003). Analyzing longitudinal data involves a deep understanding of both statistical modeling and the substantive construct of interest. In this way, modeling longitudinal data is part art, part science.

There are a wide variety of longitudinal models that focus on different aspects of the longitudinal or change process and make very different assumptions about the underlying mechanisms that influence the stability or change in data across time. Therefore, “when thinking about any repeated measures analysis it is best to first ask, what is your model for change?” (McArdle, 2009, p. 579). To determine the correct model for the analysis of longitudinal data, the researcher must have a substantive theory about whether and how the data should change over time, as well as some understanding of how observations across time should relate to each other. In addition, the purpose and the focus of the analysis and the nature of the research questions help to determine the correct longitudinal model.

Two of the most common families of models for examining longitudinal data in the social sciences are autoregressive or panel models (also sometimes referred to as Markov chain models) and individual growth curve models. Autoregressive panel models seek to describe interindividual differences in change across time, growth models focus on understanding within-person change across time (Little, 2013). There are several major differences between autoregressive panel models and growth models. First, autoregressive models estimate fixed parameters to explain the nature of change over time. In contrast, growth models allow for variance in the parameter estimates, which means that different individuals can have different expected growth trajectories. Second, autoregressive models are generally less concerned with the estimation of means or changes in level across time.

---

D. B. McCoach (✉) · H. Yu  
University of Connecticut, Storrs, CT, USA  
e-mail: [betsy.mccoach@uconn.edu](mailto:betsy.mccoach@uconn.edu)

H. Yu  
e-mail: [huihui.yu@uconn.edu](mailto:huihui.yu@uconn.edu)

Both autoregressive models and growth models allow for correlations of observations across time, but the assumed correlational structure differs between the two families of models. In autoregressive models, “a variable is expressed as an additive function of its immediately preceding value plus a random disturbance” (Bollen & Curran, 2006, p. 208). The correlational structure of an autoregressive model fits a simplex pattern, where the correlations between the adjacent time points are strongest, and the correlations between time points that are further and further apart become increasingly small. The autoregressive parameter captures the degree of stability across adjacent time points: the larger the autoregression parameter, the longer it takes for the correlations of nonadjacent time points to dampen or dissipate. The model-implied correlational structure of a growth model also allows for dependency across time; however, the structure of the errors makes different assumptions about the dependency and models the dependency using the variance components.

Not all longitudinal models involve systematic growth or decline over time. For example, imagine that a researcher collects mood data on adults every day for 3 months. Although these data are longitudinal and the researcher would expect to see day-to-day changes in mood, he or she would probably not expect to see any “growth” in mood across time. Instead, mood at any given time may be predicted by a person’s overall mean mood and some amount of random daily fluctuation or error. If mood on the prior day predicts today’s mood, then the model has an autoregressive quality. Still, such a dataset does not actually involve growth or decay over time.

In this chapter, we focus on one specific type of longitudinal model that has become quite popular in the research literature over the past decade: the individual growth model. In reading fluency research, growth curve modeling has been used to examine general growth patterns in oral reading fluency (ORF) as well as to examine differences in growth related to language impairments, disability status, English language learning status, socioeconomic status and other salient factors (Crowe, Connor, & Petscher, 2009; Logan & Petscher, 2010; McCoach, O’Connell, Reis, & Levitt, 2006; Puranik, Petscher, Al Otaiba, Catts, & Lonigan, 2008; Speece & Ritchey, 2005). Researchers can fit growth curve models to explicitly estimate systematic change in the outcome variable across time. Individual growth models have widespread appeal to developmental and behavioral researchers because they allow for the estimation of systematic growth or decline over time. Growth curve models can be estimated using either multilevel (mixed) or structural equation models (SEM). Although most basic growth models can fit in either framework, each of the techniques provides certain advantages. Therefore, in this chapter, we provide a brief introduction to individual growth curve modeling in both the multilevel and structural equation modeling frameworks. It is important to remember, however, that traditional growth curve models represent only one of many families of models that can be applied to longitudinal data.

## Purpose of Growth Curve Models

As the name suggests, growth curve models provide researchers with a method to investigate systematic growth (or decline) in outcomes across time. One of the most appealing features of individual growth curve models is that they allow for the estimation of growth parameters (such as the initial status and the growth rate) for each individual in the study. Therefore, we can estimate distinct model-implied growth rates for each individual in the study. We can also partition variance into within- and between-person variance, which allows us to better understand the nature of the change process. If most of the variability in the data is within-person variability, then change across time is responsible for the lion's share of the variability. If between-person variance represents a sizeable proportion of the total variance, then people are systematically different from each other as well as changing across time.

The simplest growth curve models assume a linear change trajectory; however, it is possible to estimate a wide variety of nonlinear trajectories by altering the specifications of the growth model. In this chapter, we demonstrate one method to estimate nonlinear growth trajectories by specifying a piecewise linear growth model. We also extend the growth framework to model multiple outcome variables. Finally, we provide a brief example of multilevel growth modeling, in which people are clustered within organizations such as schools.

## Data Requirements and Assumptions of Growth Curve Modeling

In general, the study of change requires data collected from the same units across multiple time points. Further, growth modeling techniques also require at least three waves of data to estimate simple linear trajectories. The estimation of nonlinear growth trajectories requires additional observations across time. With larger numbers of time points, it is possible to fit increasingly complex growth functions. Therefore, it is advantageous to hypothesize the nature and shape of the growth trajectory prior to collecting data to ensure that the hypothesized trajectory will be estimable. Both the number and the spacing of data collection points influence our ability to accurately capture change across time. When data points are too infrequent or when there are too few data points, it may not be possible to accurately model the functional form of the change process.

There are two additional requirements for conducting individual growth modeling. First, it is essential to document how much time has elapsed between data collection points. Fortunately, the data need not be collected across equally spaced intervals; however, we must know the length of the time interval between data collection points (Singer & Willett, 2003). When plotting growth trajectories, time is plotted on the *x*-axis and the score and the outcome variables are plotted on the *y*-axis. Therefore, knowing the distance between testing occasions allows us to plot

the dependent variable or the “*y*” score, on the correct location of the *x*-axis to correctly model the functional form of the growth.

The second requirement is that the outcome measure must be psychometrically sound and produce scores that are comparable over time (Singer & Willett, 2003). The measurement scale must also remain consistent across time. In other words, a person whose outcome score has not changed across time would receive the same score at each measurement occasion. This generally requires the use of the same assessment at multiple time points or the use of vertically scaled assessments (Singer & Willett, 2003). Vertically scaled assessments have undergone a procedure to place scores from a variety of ages or developmental levels on the same scale or metric, which allows for the direct comparison of scores over time. Subsequently, they are useful for modeling growth across time for constructs that cannot be measured using the same assessment across multiple time points. For instance, cognitive constructs such as academic achievement and cognitive ability cannot be measured using the same assessment at different stages of development. A test that measures a typical 5-year old’s reading ability would not be an appropriate assessment of a typical high school student’s reading ability and vice versa. Therefore, tests that are designed to measure the same construct at different developmental levels must be “equated” so that the scores from different sets of items given at different ages are comparable.

In the absence of vertical scaling, the difference between the two scores on two different tests does not measure growth in any meaningful way because the two scores are on two different, unlinked scales. Many academic tests are scaled within a specific content area and grade level; however, they are not designed to place scores along the same metric across time points. In such a scenario, comparing students’ scores across time cannot provide information on student growth. In addition to having a scale that provides a common metric across time, the validity of the assessment must remain consistent across multiple administrations of the assessment (Singer & Willett, 2003). For example, an ORF probe may provide a good indication of a first grader’s reading ability, but would likely not provide a good indication of a tenth grader’s reading ability.

It is possible to model dichotomous, ordinal, Poisson, or other types of non-normal outcome variables using growth modeling techniques; however, standard growth models assume that the outcome variables are continuous and normally distributed. Given that participants are changing across time, the score distribution of the outcome variable may change over time, especially when using the same measure across time. ORF, as measured by the dynamic indicators of basic early literacy skills (DIBELS), provides an excellent example in which the shape of the distribution changes across time (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009). Catts et al. found that early DIBELS scores were strongly positively skewed and exhibited strong floor effects during the first few administrations. For example, ORF was strongly positively skewed and exhibited strong floor effects though fall of first grade. By April of second grade, ORF scores were relatively normally distributed. Floor effects are particularly troubling for researchers. When many students score at the floor of the test, it is impossible to distinguish their skill levels.

Thus, in spite of the variability among those examinees in terms of their reading ability, the test is unable to detect those differences. In contrast, when tests of ORF are administered to upper elementary and middle school students, the assessments exhibit substantial ceiling effects and the scores on the ORF are negatively skewed. By late elementary school, the ORF passages would be quite easy for most of the students. Further, there is an upper limit on the speed at which a person can read aloud. Therefore, the distribution of scores for older children and adults tends to be negatively skewed: Few examinees have low scores and many students are able to reach a maximum possible score of words read aloud correctly in 1 min. The data in this chapter exhibit such properties: The ORF distribution of ORF scores is positively skewed at the first data collection point, which occurs in the fall of first grade, but it becomes more normal by the eighth data collection point, which occurs at the end of second grade.

When the shape of the distribution changes across time, transformation of the outcome variable is not a viable option. Transforming the dependent variable in the same manner across all of the time points would alleviate the problem at some time points but exacerbate the issue at another time points. Applying different transformations across the different time points is also not possible as it would place the outcome variable on different and incomparable scales across time. Standardizing the dependent variable at each time point to normalize the distribution and to try to ensure that the scores are equated across time is also a poor idea. Recall that standardized scores have a mean of 0 and a standard deviation of 1 (and thus a variance of 1). Therefore, the mean at every time point is 0. Assume that children's ORF scores do increase across time. A child with a standardized score of 0 at every time point would actually be growing. The score of 0 tells us that the student is scoring at the 50th percentile on reading at each time point. Because the mean is standardized to be 0 for each of the time points, growth models using standardized scores are not capturing growth per se, instead, they capture change in relative status across time. Second and more importantly, standardizing scores at each time point constrains the variance of the measure to be equal across time. Yet when scores are systematically changing across time, we would not expect the variance to be constant across time. In fact, if the variance remains constant across time and students' scores increase, then the correlation between students' initial status and their growth rate must be negative (Campbell & Kenny, 1999). The variance in achievement, skills, or ability generally increases across time (Bast & Reitsma, 1998; Gagné, Wager, Golas, Keller, & Russell 2005; Kenny, 1974), and there is generally a positive correlation between students' initial reading scores and their reading growth across time. In fact, the observation of Matthew effects in reading is not possible unless the variance in reading scores increases across time. Therefore, standardizing the variable of interest to have a constant mean and a constant variance at each time point "constitutes a completely artificial and unrealistic restructuring of interindividual heterogeneity in growth" (Willett, 1989), which is likely to produce distorted results (Thorndike, 1966; Willett, 1989). Thus, it is inadvisable to standardize the dependent variable when conducting growth curve analyses (Willett, 1989).

## **Time-Structured and Time-Unstructured Data**

Data are time structured if all the units are measured on the same data collection schedule (i.e., at the same time points). The time-structured nature requires that the interval between data collection points 1 and 2 and the interval between data collection points 2 and 3 is equal across all students in the sample (Kline, 2005). In contrast, when time intervals can vary both within and across people, such data are often referred to as “time-unstructured” data (Singer & Willett, 2003). Time-unstructured data are collected on different schedules or at different time points (Skrondal & Rabe-Hesketh, 2008) and the data collection schedule can be completely different for every person in the sample. When using multilevel analyses, each participant can have their own unique data collection schedule. In other words, each person could have a different number of observations spaced across different time intervals.

Growth curve models are more flexible than traditional repeated measures analyses in that they can accommodate missing data, as long as the data are missing at random. With time-unstructured data, we can flexibly model any data collection schedule. This flexibility allows us to easily model growth trajectories for participants with missing data.

## **Model Specification for Basic Growth Curve Models**

In the following sections, we describe the data structure and model specification for multilevel growth models. Afterward, we describe the analogous structural equation modeling specifications for growth models.

### ***Data Structure for Growth Curve Models***

Individual growth analysis requires that individuals are measured repeatedly on the same outcome variable. Therefore, the repeated measures are nested within individuals and they are correlated with each other across time. Because multilevel approaches to growth curve modeling view observations across time as nested within people, and because time enters the model as an explicit independent variable, multilevel approaches to growth curve analyses require long, univariate, person–period data files. In such a structure, each observation at a given time point becomes a row in the data file. Therefore, each unit (or person) occupies multiple rows within the person–period data file (Singer & Willett, 2003). Generally, data are stored in a multivariate or wide file, where each row denotes a separate person. Further, when using SEM to fit growth models, the data need to be arranged in a wide file. Still, it is quite simple to restructure data from a wide file to a long file to use multilevel techniques. Figure 10.1 depicts the data restructure wizard in statistical package for the social sciences (SPSS). This tool seamlessly restructures data from a wide format to a long format or from a long format to a wide format. Most major software



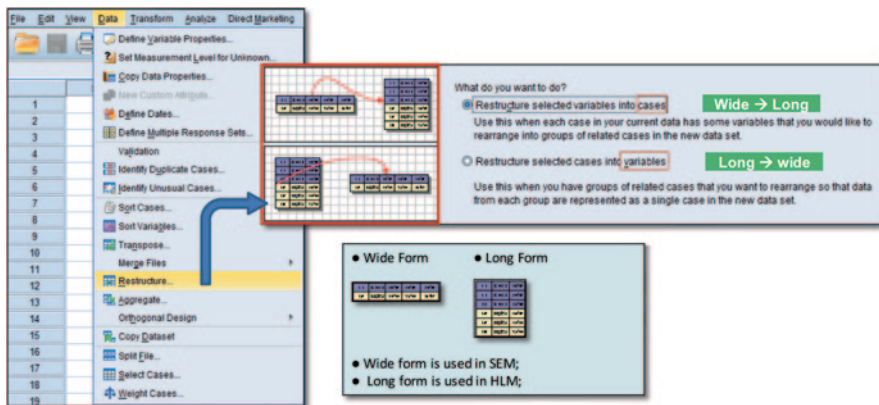


Fig. 10.1 SPSS data restructure wizard, which converts wide files to long files

programs have a similar tool or command that allows for the convenient restructuring of data into a long, person–period dataset.

### *The Two-Level Multilevel Model for Linear Growth*

From a multilevel perspective, a simple growth curve model has two levels: a within-individual level (level-1) and a between-individual level (level-2). Observations across time are the level-1 units, and they are nested within people, who are the level-2 units. Thus, the level-1 model captures the shape of an individual’s growth trajectory across time and includes any covariates that vary across time within individual. In other words, the variable can take on a different value for the same person at different time points throughout the study period. In a study of reading fluency, time-varying covariates might include scores on other language or reading assessments, the number of instructional minutes that the student spends engaged in reading activities, or whether or not the student received additional support services during the time period in question. The level-2 model captures the between-person variability in the growth parameters across time and includes any covariates that vary across individuals but are constant across time within a given individual. Examples of level-2 covariates include gender, race/ethnicity, or other stable, personal characteristics. The equations below depict a simple two-level linear growth model.

Level-1:

$$y_{ii} = \pi_{0i} + \pi_{1i}(time_{ii}) + e_{ii}.$$

Level-2:

$$\pi_{0i} = \beta_{00} + r_{0i}.$$

$$\pi_{1i} = \beta_{10} + r_{1i}.$$

The dependent variable ( $y_{it}$ ) is the score for student  $i$  at time  $t$ , which is a function of the intercept,  $\pi_{0i}$ , (which is the predicted value of  $y_{it}$  when time=0),  $\pi_{1i}(\text{time}_{it})$ , and individual error. The time slope,  $\pi_{1i}$ , represents the linear rate of change over time. As we mentioned earlier, because time enters the equation as an independent variable, each participant can have its own unique values on the time variable. To make the intercept interpretable, the time variable is generally centered on some meaningful value that occurs during the data collection period. The most common approach is to center time at the beginning of the study period. Using such a centering scheme, the intercept represents the expected value at the beginning of the study. If age is used as the time variable, it is quite common to center at a particular age (for example, at age 6). This has the added advantage of controlling for age in addition to centering time. For example, if age/time=0 occurs when each of the students is 6 years old, then the intercept represents the model predicted score at age 6.

Notice that both the slope and the intercept contain a subscript  $i$ , indicating that a separate slope and intercept are estimated for each person in the sample. The deviation of a particular observation from the model-predicted trajectory is captured in the error term, ( $e_{it}$ ), which represents the within-person measurement error associated with that individual's data at that time point. The pooled error variability within individuals' trajectories is estimated by the variance of  $e_{it}$  [ $\text{var}(e_{it}) = \sigma^2$ ] (Raudenbush & Bryk, 2002), and this error variance is generally assumed to be constant across time.

The level-2 equation models the average growth trajectory across people. It also captures between-person differences in the model-implied growth trajectories based on level-2 (time invariant) covariates. The second level of the multilevel model specifies that the randomly varying intercept ( $\pi_{0i}$ ) for each individual ( $i$ ) is predicted by an overall intercept ( $\beta_{00}$ ), the effects of any level-2 predictors on the intercept, and  $r_{0i}$ , the level-2 residual, which represents the difference between person  $i$ 's model-predicted intercept (based on the overall intercept,  $\beta_{00}$ , and level-2 predictors) and his or her actual intercept. Likewise, the randomly varying linear growth slope ( $\pi_{1i}$ ) for each individual ( $i$ ) is predicted by an overall intercept ( $\beta_{10}$ ), the effects of level-2 variables on the linear growth slope, and  $r_{1i}$ , the level-2 residual, which represents the difference between person  $i$ 's model-predicted linear growth slope and his or her actual growth slope. Imagine a simple scenario in which time is coded 0, 1, 2. The intercept,  $\pi_{0i}$ , represents the predicted initial status of person  $i$ . Thus,  $\beta_{00}$  represents the overall average intercept. The linear growth parameter ( $\pi_{1i}$ ) represents the average growth rate ( $\beta_{10}$ ). The between-student variability in the intercept is captured by the variance of  $r_{0i}$  [ $\text{var}(r_{0i}) = \tau_{00}$ ]. If the intercept is centered on initial status, then  $\tau_{00}$  represents the between-person variability in initial status, or where people start. Likewise, the amount of variability in the time slope between students is estimated by the variance of  $r_{1i}$  [ $\text{var}(r_{1i}) = \tau_{11}$ ] (Raudenbush & Bryk, 2002):  $\tau_{11}$  represents the between-person variability in peoples' growth rates. The inclusion of the  $r_{0i}$  and  $r_{1i}$  in the level-2 equations allows for between-person variability in the intercepts and slopes. In addition, we generally estimate their covariance,  $\tau_{01}$ . The standardized  $\tau_{01}$  estimate from the model above (which does not yet include any level-2 predictors) provides the correlation between initial status (or, more generally, the intercept) and growth.

Generally, we start by estimating the model above and we develop the final level-1 model prior to including person-level (level-2) predictors of the level-1 parameters. To develop the level-1 model, first we consider the shape of the growth trajectory and try to fit a model that adequately captures the shape of the growth trajectory without any time-varying covariates (which are level-1 predictors that vary across time). Next, we introduce time-varying covariates into the level-1 model. Finally, we consider any potential interactions among time variables and time-varying covariates. After we are satisfied with the level-1 model, we turn our attention to the level-2 model. At this point, we can include person-specific variables in the level-2 model to explain variation in the intercept and the growth slope. For instance, person-level predictors such as gender, student's socioeconomic status, or their cognitive ability may help to explain where students start and how fast they grow in terms of their reading fluency. Ideally, if person-level covariates help to explain some of the interindividual variability in terms of where people start (the intercept) or how fast people grow (the slope), then the variances for  $r_{0i}$  and  $r_{1i}$  should decrease as those predictors are included in the model. Once level-2 predictors are included in the model,  $\tau_{00}$  becomes the residual variance in the intercept after controlling for the covariates. Likewise,  $\tau_{11}$  becomes the amount of between-student variability in the time slope after accounting for the person-level covariates (Raudenbush & Bryk, 2002). The standardized  $\tau_{01}$  estimate from the conditional model above is the residualized correlation between initial status (or, more generally, the intercept) and growth, which provides the relation between initial status and rate of change after controlling for the other variables in the model.

### ***Piecewise Growth Models***

Often, growth trajectories may not be modeled well by a single linear slope or rate of change, even after adjusting for time-varying covariates. There may be scenarios in which a growth pattern might be more aptly represented by dividing the trajectory into growth segments corresponding to fundamentally different patterns of change (Collins, 2006). For example, imagine that a reading researcher collects achievement data on elementary students across an entire calendar year. Time points between September and June capture the span of time for the change in achievement across the school year, whereas the period between June and the end of August captures the span of time for the change in reading scores during the summer (noninstructional) months. The achievement slope is likely to be substantially steeper and constant during instructional months and flatter (or perhaps even negative) during the summer, when students receive no academic instruction: a single linear growth parameter cannot represent the data well in this situation. Piecewise linear growth models “break up the growth trajectories into separate linear components” (Raudenbush & Bryk 2002, p. 178), and can be particularly valuable when comparison of growth rates between the separate components are of interest. Piecewise models also allow researchers to investigate differences in substantive predictors of growth between the components. Note that a sufficient number of timepoints are required to enable modeling of a separate slope for each component.

Piecewise regression techniques conveniently allow for changes in a linear growth slope across time. To achieve these representations, we include multiple-time variables into the model to capture the multiple linear growth slopes. For example, if we expect one rate of growth for time points 1–4, and another rate of growth for time points 4–8, we would introduce two time variables. The second time variable always clocks the passage of time, starting (from 0) at the point at which the discontinuity or change in slope is expected. Following our above example, our two-piece linear growth model would then be expressed as follows:

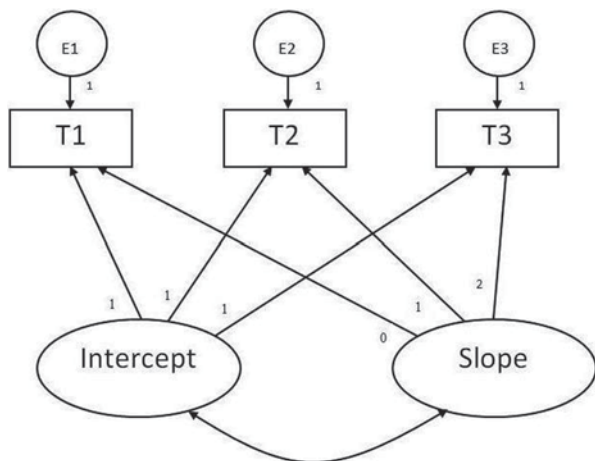
$$\begin{aligned}
 Y_{ii} &= \pi_{0i} + \pi_{1i}(\text{time\_piece1})_{ii} + \pi_{2i}(\text{time\_piece2})_{ii} + e_{ii} \\
 \pi_{0i} &= \beta_{00} + r_{0i} \\
 \pi_{1i} &= \beta_{10} + r_{1i} \\
 \pi_{2i} &= \beta_{20} + r_{2i}.
 \end{aligned}$$

Generally, reading scores exhibit substantial growth during the school year; however, reading scores often remain flat (or even decrease) over the summer months. Therefore, to adequately capture the growth in ORF, we need to fit at least two linear growth trajectories: one for school-year growth and another for summer growth. To model these multiple trajectories, we can create two time variables: one that clocks the passage of time from the beginning of the study that occurs during the school year (*time\_piece1*) and another that clocks the passage of time during the summer (*time\_piece2*). If we could assume that school-year growth remained constant within child across the multiple years of the study and summer growth also remained constant within child across the study, we could capture the zig-zag pattern of growth across the multiple years of the study with only two different slope parameters:  $\beta_{10}$ , which would capture the school-year slope, and  $\beta_{20}$  which would capture the summer slope. Thus, creative use of piecewise regression models can capture a variety of patterns of nonlinear change as well as discontinuities in growth. Of course, multiple changes in linear growth rates can be captured through piecewise models as well. In our current dataset, reading growth is measured at eight time points: four times a year across two school years. We model such data using a three-piece growth model. The first slope captures change across the first-grade year, the second slope captures the summer growth rate, and the third slope captures the growth rate during the second-grade year. Although the piecewise model allows us to flexibly model two or more growth rates that occur across particular time intervals, they are just one type of nonlinear model. Often polynomial models or other types of nonlinear models are used to capture nonlinear trajectories. The interested reader should consult O’Connell et al. (2013) for further details on specifying and interpreting polynomial growth models.

### ***SEM Models***

The SEM approach to growth curve analysis formulates individual growth curves (the intercept and the linear growth rate) as latent variables. Therefore, the mean

**Fig. 10.2** Graphic representing the estimation of a linear growth model from data collected at three time points using SEM



and the variance of the latent variables provide the parameter estimates of interest for the individual growth curves. The means of the latent variables provide the expected values for the intercept and slope parameters. In other words, the mean of the mean factor depicts the average level within the sample and the mean of the slope factor describes the average rate of change across time. The variances of the latent variables capture the between-person variance in the slope and the intercept. If the slope variance were zero, then everyone in the sample would be growing at the same rate, a common assumption in repeated measures analysis of variance (ANOVA). In traditional SEM, time is introduced through the factor loadings for the latent variable representing the linear-growth slope (Stoel & Garre, 2011); therefore, standard SEM approaches to growth curve modeling require time-structured data.

Figure 10.2 provides a graphical depiction of a linear growth model with data collected at three time points, modeled within the SEM framework. This model contains two latent variables: the slope and the intercept. To model growth using SEM requires the estimation of a model that includes means (and intercepts) as well as variances and covariances. The means of the latent variables provide the parameter estimates for the expected value of the slope and the intercept. The variances of these latent variables provide information about the between-person variability in the intercept and the slope. The factor loadings for the intercept factor are fixed to be 1 across time. The factor loadings for the slope variable contain the information about the passage of time. If we want to center time at the time of the first observation, then we would constrain the path from the slope factor to T1 (which is the outcome variable at time 1) to 0. Then, the mean of the intercept factor represents the expected value of the outcome variable at time 1. The path coefficient for the path from the slope factor to T2 represents the amount of time that has elapsed since the first data collection at time 1. Here, the path is fixed at 1 because 1 year has elapsed since the collection of T1 data. The third path is fixed to 2, which represents the total amount of time that has elapsed since the first data collection (T1). In other words, the model above illustrates a model in which the outcome variable is measured at three time points, each of which are spaced 1 year apart. Therefore, the mean of the slope

represents the yearly growth rate because the fixed path coefficients are measured in the metric of “years.” The SEM approach is analogous to the multilevel approach. But because the information about the passage of time is contained in the path coefficients instead of in a separate variable, everyone must have the same basic data collection schedule. This is why traditional SEM analyses require time-structured data whereas multilevel models can accommodate time-unstructured data.

The residual variances at different time points represent the time-specific measurement error in the outcome variable (Bollen & Curran, 2006; Muthén & Muthén, 1998). Conventional multilevel models pool this measurement error across time points and people. Therefore, in the most traditional or standard growth models all of the error variances are constrained to be equal. This traditional growth model estimates six parameters: two latent means, two latent variances, the covariance between the two latent variables, and one error variance. To identify the mean structure of the model, we set the intercepts for the outcome variable to be 0 across all time points. Conceptually, this means that the intercept and the growth slope explain change in the level of the observed scores across time. It is not uncommon to allow the variances of these residual terms to vary across time in SEM. Duncan et al. (2006) provide a readable introduction to growth curve models in SEM and Bollen and Curran (2006) provide a more advanced treatment of the topic.

### ***Multilevel Models versus Structural Equation Models***

Traditional SEM latent growth models require time-structured data; this represents a major difference between the SEM and hierarchical linear modeling (HLM) approaches to individual growth modeling, and is one major advantage to modeling growth data within a multilevel framework. Every participant can have their own unique data collection schedule. Thus, traditional multilevel models are inherently flexible in modeling unstructured data. In contrast, because the information about time is contained in the factor loadings, traditional SEM require time-structured data. If data are time unstructured and preserving that structure is important, multilevel approaches to growth curve analysis provide a seamless method of modeling time in a purely unstructured fashion.

But there are several advantages to using the SEM approach to growth curve modeling. First, whereas multilevel models provide a variety of a priori error covariance structures within the mixed-model framework, it is easier to specify a wide variety of “ad hoc” error covariance structures within the SEM framework. Second, SEM provides the ability to build a measurement model for the outcome variable, allowing for the incorporation of multiple indicators at each time point. In other words, the outcome variable itself can be a latent variable. For example, instead of using one measure of ORF as an outcome variable, we can create a latent variable called fluency, which is comprised of multiple (preferably three or more) ORF trials. Utilizing this latent variable strategy allows researchers to disaggregate measure-specific error from time-specific error, which is a tremendous benefit of using SEM (Kline & ebrary, 2011). Third, SEM models can incorporate latent variables as predictors. Accounting for measurement error in the predictors should result in less-biased parameter



estimates. Fourth, SEM models can accommodate mediational models far more easily than multilevel models can (Little, 2013). Fifth, SEM models allow for the simultaneous or sequential estimation of multiple growth models for multiple-dependent variables (Duncan, Duncan, & Strycker, 2006). Sixth, SEM allows for more flexibility in the incorporation of time-varying covariates into the growth model (Curran, Lee, Howard, Lane, & MacCallum, 2012). Finally, because the information about time is contained in the factor loadings, when there are at least four time points it is possible to fix two factor loadings and free the rest of the loadings, allowing for a very flexible expression of the growth trajectory (Bollen & Curran, 2006). Therefore, SEM provides a great deal of flexibility for modeling growth. The SEM and multilevel approaches can be combined, using multilevel SEM, which can capture the nested nature of educational data and capitalize on the strengths of the SEM tradition.

Growth trajectories can be influenced by many factors. These factors may be observed or latent and time varying or time invariant. If these factors can be measured quantitatively, then they could be included as exogenous variables in growth curve models to better explain the nature of the growth process. Because our focus is to demonstrate how to model a growth trajectory, we do not use any time-varying or time-invariant covariates in our models in this chapter. When including covariates, we recommend the following model-building process: First, include only the time variables, next include any other level-1 variables, and, as a final step, add level-2 predictors. We now turn our attention to an applied example to illustrate the utility of growth curve modeling within fluency research.

## Applied Example

To illustrate the use of individual growth curves to model fluency data, we use two DIBELS measures: nonword fluency (NWF) and ORF. We begin with a series of univariate analyses using ORF data. We contrast the multilevel and SEM approaches to growth curve analysis, fitting the same growth models to the data under the two approaches. In addition, we compare the fit of linear and piecewise growth models to these data. Lastly, we introduce NWF into our analyses to demonstrate how to model multiple outcome variables in one model. We use the HLM 7 program to fit the multilevel models and Mplus 7 to fit the SEM models.

### *Step 1: Understanding Our Data*

Our dataset contains 18,667 students, measured quarterly at eight time points across kindergarten and first grade. A full description of the data can be found in Chap. 12 (Petscher, Koon, Kershaw) of this volume. We cannot overstate the importance of visually inspecting both individual growth trajectories and a plot of the change in the means on the outcome variable across time. No modeling technique, no matter how novel or sophisticated, can substitute for a solid understanding of nature of the observed change within the dataset (McCoach et al., 2013). Prior to conducting



any statistical analysis, we examined the raw data to describe the observed growth trajectories. We present the means, standard deviations (SD), and correlations for the eight repeated measures of NWF and ORF in Table 10.1. Our preliminary descriptive analyses revealed several points of interest. First, the correlations among the repeated measures of the same construct are moderately high (ranging from .55 to .81 for NWF and ranging from .64 to .94 for ORF) and the correlations among the repeated measures across constructs are also moderately high (ranging from .49 to .76). Observations of the same measure over time are clearly associated, and the two measures, NWF and ORF, are also correlated with each other. Second, the magnitudes of the correlations decrease as the time intervals between observations increase. Figure 10.3 plots the correlations within and across constructs over time. Third, as expected, the means of ORF and NWF increase over time (Table 10.1). Yet the growth rates are not consistent over time. Thus, a linear model does not seem appropriate to capture the growth in reading fluency across this 2-year period. The ORF and NWF slopes are markedly less positive between time point 4 (ORF 4 and NWF4) and time point 5 (ORF 5 and NWF5), which encompass the summer vacation period between first grade and the second grade. According to the plots in Fig. 10.4, students make positive growth during the first four data collection points (which are collected during first grade). Time 4 is the last data collection in first grade. Time 5 is the first data collection point in grade 2. Therefore, the summer break occurs between data points 4 and 5. During the summer (which is captured by examining the line between data points 4 and 5), students regress slightly in NWF and make little to no progress in ORF. Given this even pattern of development, we would expect a nonlinear model such as a piecewise model to fit the data better than a simple linear model. Lastly, the variances in the repeated measures increase over time. Such a pattern is common in reading research, and it indicates that students are becoming increasingly more different from each other (Bast & Reitsma, 1997, 1998; McNamara, Scissons, & Gutknecht, 2011).

To demonstrate univariate growth models, we present a series of growth models using the ORF data. The NWF fluency data look quite similar; therefore we do not repeat the process with NWF. To demonstrate simple multivariate growth models, we utilize both NWF and ORF.

### ***Model 1: The Unconditional Single-Process Growth Curve Model***

First, we model growth in ORF across the eight time points. According to the plot (b) in Fig. 10.4, the growth trajectories change across time; there is not one constant growth rate. Instead, there is one elbow at time point 4 and another one at time point 5. In addition, the growth between these time points is substantially lower than the growth between time points 1–4 and time points 5–8.

For comparison, we present two unconditional models: a linear growth model and a piecewise growth curve model. The linear growth model fits one slope across all eight time points, thus assuming that the growth is consistent over time. But given the shape of the observed data, such a model is unlikely to provide the best fit to

**Table 10.1** Descriptive statistics and correlations among the repeated measures of NWF and ORF

	NWF-1	NWF-2	NWF-3	NWF-4	NWF-5	NWF-6	NWF-7	NWF-8	ORF-1	ORF-2	ORF-3	ORF-4	ORF-5	ORF-6	ORF-7	ORF-8
NWF-1	1.00															
NWF-2	.76	1.00														
NWF-3	.72	.77	1.00													
NWF-4	.66	.71	.78	1.00												
NWF-5	.65	.69	.74	.77	1.00											
NWF-6	.60	.65	.70	.75	.79	1.00										
NWF-7	.59	.62	.68	.73	.76	.81	1.00									
NWF-8	.55	.58	.63	.68	.72	.76	.81	1.00								
ORF-1	.76	.67	.64	.58	.58	.52	.52	.48	1.00							
ORF-2	.75	.72	.70	.65	.63	.58	.58	.55	.91	1.00						
ORF-3	.73	.70	.74	.69	.67	.62	.62	.60	.86	.94	1.00					
ORF-4	.70	.69	.72	.72	.70	.65	.65	.63	.80	.89	.93	1.00				
ORF-5	.68	.66	.69	.68	.72	.64	.65	.62	.77	.85	.89	.92	1.00			
ORF-6	.66	.65	.68	.67	.69	.67	.67	.65	.72	.81	.86	.90	.93	1.00		
ORF-7	.63	.62	.65	.66	.68	.65	.70	.66	.67	.76	.82	.87	.90	.93	1.00	
ORF-8	.61	.61	.64	.64	.66	.64	.66	.67	.64	.73	.79	.84	.87	.91	.93	1.00
Mean	37.02	52.51	58.65	69.06	65.95	80.92	87.09	95.20	21.08	32.91	42.38	52.65	56.09	69.18	81.70	92.70
S.D.	21.62	24.57	28.06	31.88	31.66	37.20	38.62	42.38	19.40	25.23	30.69	30.54	31.67	31.82	34.58	35.25

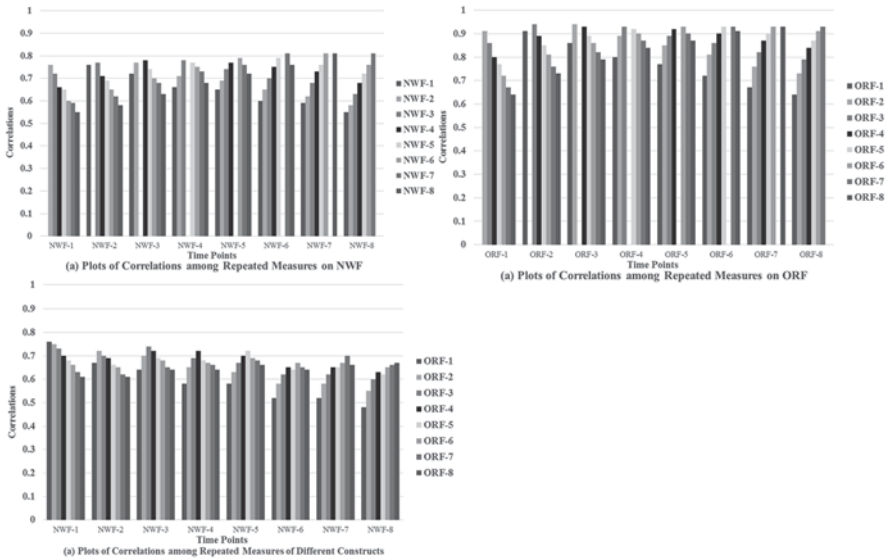


Fig. 10.3 The Plots of the correlations among the repeated measures

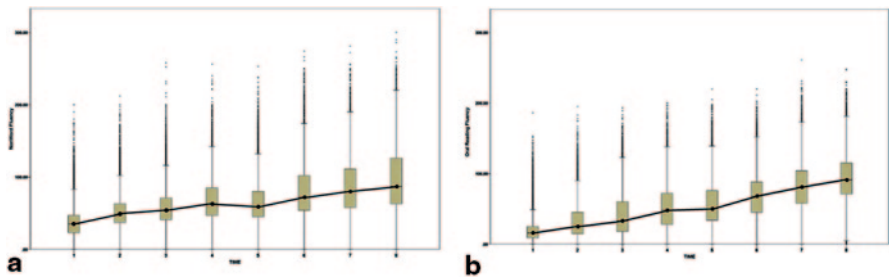


Fig. 10.4 The Boxplots of the dependent variables

the data. As described earlier, a piecewise linear model consists of multiple pieces, each of which are linear slopes. Although all of the pieces are themselves linear, the resulting growth trajectory is a nonlinear amalgam of two or more linear segments.

For the piecewise growth curve model, we fit three linear pieces (i.e., slopes). The first slope represents the growth rate from time 1 to 4; the second slope represents the growth rate from time 4 to 5; and the third slope represents the growth rate from time 5 to 8. Assuming that data are time structured, Table 10.2 presents the design matrix for these two models.  $X_{0tj}$  represents the multiplier or factor loading for the intercept and it is always 1. In the linear model,  $X_{1tj}$  represents the unit change in time across all eight time periods. In the piecewise linear model,  $X_{1tj}$  represents the unit change in time between time points 1 and 4,  $X_{2tj}$  represents the unit change in time between time points 4 and 5, and  $X_{3tj}$  represents the unit change in time

**Table 10.2** Examples of the arbitrary values of the latent slopes' loadings

Time (t)	Linear growth model ( $n=1$ )		Piecewise growth model ( $n=3$ )			
	$x_{0tj}$	$x_{1tj}$	$x_{0tj}$	$x_{1tj}$	$x_{2tj}$	$x_{3tj}$
1	1	0	1	0	0	0
2	1	1	1	1	0	0
3	1	2	1	2	0	0
4	1	3	1	3	0	0
5	1	4	1	3	1	0
6	1	5	1	3	1	1
7	1	6	1	3	1	2
8	1	7	1	3	1	3

between time points 5 and 8. To appropriately model a piecewise growth trajectory with separate slopes, only one time variable changes from (t-1) to t.

If students grew at the same rate across first and second grades, we could easily simplify the model above to fit a two-piece model in which there are only two rates of change: Rate 1 would capture the growth rates from times 1-4 and times 5-8 and rate 2 would capture the summer growth (from times 4-5). In the current example, such a two-piece model does not fit the data as well as the three-piece model does but, in other scenarios or datasets, it certainly could.

We used both multilevel modeling and SEM to analyze these data, using both linear and piecewise linear models. Table 10.3 presents these two approaches side by side for comparison. The initial SEM models constrain all the residual variances to be equal across time for direct comparison to the traditional HLM approach. We also provide results from the SEM models that allow for heterogeneous error variances across time.

Table 10.4 summarizes the variance components and residual variances for the linear growth model; Table 10.5 summarizes the variance components and residual variances of the piecewise growth model. For both linear growth models and piecewise growth models after controlling for the growth factors, the residual variances do differ across the eight time points ( $p < .0001$ ) and the fit is slightly better for these models (The BIC for the piecewise model is 1145099.2, whereas the BIC for the linear model is 1166288.2). But constraining the within-person residual variance to be equal across all eight time points has a fairly trivial influence on the estimated between-person residual variances for the intercept and growth slopes and no appreciable effect on the estimates of the intercept and growth slopes.

Comparing the constrained linear growth model with the constrained piecewise growth model, the within-student across-time variance ( $\sigma^2$ ) from the linear growth model is 90, and the within-student across-time variance ( $\sigma^2$ ) decreases to 65. In other words, using the three piecewise growth models helps to explain  $(90 - 65)/90 = 25/90 = 27.8\%$  more of the within-student (across time) variance than the standard linear growth model does.

Table 10.6 summarizes the estimated fixed effects obtained from the three models and compares the model-predicted values to the actual sample values for each

**Table 10.3** Comparing the multilevel modeling approach and the SEM approaches

Multilevel model	SEM model with homogeneous error variances
<p>Linear growth models</p> <p>Level-1 Model:  <math>y_{ij} = i_{0j} + s_{1j} * (x_1)_{ij} + e_{ij}</math></p> <p>Level-2 Model:  <math>i_{0j} = \beta_{00} + r_{0j}</math>  <math>s_{1j} = \beta_{10} + r_{1j}</math></p>	
<p>Default assumption:                      The level-1 residual variance within students is constrained to be equal across times, which is the default assumption in the HLM program</p> <p>Two Fixed effects:  <math>\beta_{00}</math> : the expected (mean) initial score  <math>\beta_{10}</math> : the expected (mean) growth rate</p> <p>Three Random effects:                      Level-1:                      Variance of <math>e_{ij}</math> (<math>\sigma^2</math>): the within-student (residual) variance, which is assumed homogenous across time points                      Level-2:                      Variance of <math>r_{0j}</math> (<math>\tau_{00}</math>): the between-person (residual) variance in the initial ORF scores                      Variance of <math>r_{1j}</math> (<math>\tau_{11}</math>): the between-person (residual) variance in the growth slope</p>	<p>In the SEM model, error variances of e1–e8 are constrained to be equal. This constrained SEM model is identical to the two-level HLM model on the left side</p> <p>The residual variance of the 8 observed outcome variables corresponds to the level-1 error variance (<math>\sigma^2</math>) in the HLM model</p> <p>The latent intercept variable “i” is the initial ORF level at time one, which corresponds to <math>i_{0j}</math> in the HLM model</p> <p>Therefore, the mean of “i” corresponds to the fixed effect of <math>i_{0j}(\beta_{00})</math> and the variance of “i” corresponds to the random effect of <math>i_{0j}(\tau_{00})</math></p> <p>The latent slope variable s1 is the latent growth rate, which corresponds to the fixed effect of <math>s_{1j}(\beta_{10})</math> The variance of s1 corresponds to the random effects of <math>s_{1j}(\tau_{11})</math></p>
<p>Piecewise growth models</p> <p>Level-1 Model:  <math>y_{ij} = i_{0j} + s_{1j} * (x_1)_{ij} + s_{2j} * (x_2)_{ij} + s_{3j} * (x_3)_{ij} + e_{ij}</math></p> <p>Level-2 Model:  <math>i_{0j} = \beta_{00} + r_{0j}</math>  <math>s_{1j} = \beta_{10} + r_{1j}</math>  <math>s_{2j} = \beta_{20} + r_{2j}</math>  <math>s_{3j} = \beta_{30} + r_{3j}</math></p>	<p>Error variances of e1–e8 are constrained to be equal</p>

**Table 10.3** (continued)

Multilevel model	SEM model with homogeneous error variances
<p>Four fixed effects:</p> <p><math>\beta_{00}</math> : the expected value of the initial score</p> <p><math>\beta_{10}</math> : the expected growth rate from time 1 to time 4</p> <p><math>\beta_{20}</math> : the expected growth rate from time 4 to time 5</p> <p><math>\beta_{30}</math> : the expected growth rate from time 5 to time 8</p> <p>Five random effects:</p> <p>Level-1:</p> <p>Variance of <math>e_{ij}</math> (<math>\delta^2</math>): the within-person variance which is assumed to be homogenous across time points</p> <p>Level-2:</p> <p>Variance of <math>r_{0j}</math> (<math>\tau_{00}</math>): the between-person variance in the initial ORF scores</p> <p>Variance of <math>r_{1j}</math> (<math>\tau_{11}</math>): the between-person variance in the first slope, which represents the growth rate from time 1 to time 4</p> <p>Variance of <math>r_{2j}</math> (<math>\tau_{22}</math>): the between-person variance in the second slope, which represents the growth rate from time 4 to time 5</p> <p>Variance of <math>r_{3j}</math> (<math>\tau_{33}</math>): the between-person variance in the third slope, which represents the growth rate from time 5 to time 8</p>	<p>The latent intercept variable “<math>i</math>” is the initial ORF level at time one, which corresponds to <math>i_{0j}</math> in the HLM model</p> <p>Therefore, the mean of “<math>i</math>” corresponds to the fixed effect of <math>i_{0j}(\beta_{00})</math> and the variance of “<math>i</math>” corresponds to the random effect of <math>i_{0j}(t_{00})</math></p> <p>The latent slope variables “<math>s_1</math>”, “<math>s_2</math>”, and “<math>s_3</math>” are respectively representing the growth rate from time 1 to time 4, the growth rate from time 4 to time 5, and the growth rate from time 5 to time 8. The means of “<math>s_1</math>”, “<math>s_2</math>”, and “<math>s_3</math>” correspond to the fixed effects of <math>s_{1j}(\beta_{10})</math>, <math>s_{2j}(\beta_{20})</math>, and <math>s_{3j}(\beta_{30})</math>. The variances of “<math>s_1</math>”, “<math>s_2</math>”, and “<math>s_3</math>” correspond to the random effects of <math>s_{1j}(\tau_{11})</math>, <math>s_{2j}(\tau_{22})</math>, and <math>s_{3j}(\tau_{33})</math></p>

**Table 10.4** Estimated variance components for the linear growth model

HLM		Constrained SEM		Unconstrained SEM	
$\sigma^2$	90.07	Residual variance constrained as equal	90.08	Residual variance ORF1	85.039
				Residual variance ORF2	33.086
				Residual variance ORF3	93.496
				Residual variance ORF4	92.967
				Residual variance ORF5	119.74
				Residual variance ORF6	69.117
				Residual variance ORF7	87.25
				Residual variance ORF8	129.556
$\tau_{00}$	541.55	Variance of i	541.70	Variance of i	551.31
$\tau_{11}$	12.51	Variance of s1	12.51	Variance of s1	13.48
Model fit indices					
AIC		1166241.2			1161788.5
BIC		1166288.2			1161890.4
Adjusted BIC		1166269.2			1161849.1

*BIC* bayesian information criterion, *AIC* akaike information criterion

**Table 10.5** Estimated variance components for the piecewise growth model

HLM		Constrained SEM		Unconstrained SEM	
$\sigma^2$	65.19	Residual variance constrained as equal	65.20	Residual variance ORF1	23.24
				Residual variance ORF2	47.31
				Residual variance ORF3	74.46
				Residual variance ORF4	55.74
				Residual variance ORF5	67.17
				Residual variance ORF6	64.70
				Residual variance ORF7	88.97
				Residual variance ORF8	81.91
				$\tau_{00}$	389.95
$\tau_{11}$	28.54	Variance of $s1$	28.53	Variance of $s1$	34.86
$\tau_{22}$	47.93	Variance of $s2$	47.91	Variance of $s2$	94.20
$\tau_{33}$	18.65	Variance of $s3$	18.65	Variance of $s3$	46.26
Model fit indices					
AIC		1144981.7		1142007.0	
BIC		1145099.2		1142179.3	
Adjusted BIC		1145051.5		1142109.4	

of the eight time points. Using the parameter estimates from the linear model, we predict that students' initial ORF is approximately 21 words ( $B_{00} = 21.06$ ) and that they gain approximately 10 words per benchmark assessment ( $B_{10} = 9.94$ ) across the 2-year period. Using the piecewise linear growth model, we still estimate that students' initial ORF is 21 words ( $B_{00} = 21.40$ ). But their reading fluency growth rate is approximately 10 points ( $B_{10} = 10.38$ ) per quarter during the first-grade school year, 4 points ( $B_{20} = 3.84$ ) during the summer break between first and second grade, and 12 points ( $B_{30} = 12.30$ ) per quarter during the school year of second grade. The results of the piecewise linear model indicate that reading fluency growth is faster during the school year than during the summer and that ORF grows faster in second grade than it does in first grade.

Figure 10.5 plots the trajectories for both ORF and NWF. Clearly, the data for both measures appears to be nonlinear: students make more growth on both ORF and NWF during the school year than they do in the summer.

Figure 10.6 plots both the observed means and estimated means for the linear and piecewise models. The piecewise growth models fit the sample data better than the linear growth models do, especially for time points 5 and 8. In contrast, the linear models overestimate the mean at time 5 and underestimate the mean at time 8.



**Table 10.6** Observed and estimated means at each time point

Model	Approach	Fixed effects	Mean	T1	T2	T3	T4	T5	T6	T7	T8
Linear growth model	HLM	$\beta_{00}$	21.06	1	1	1	1	1	1	1	1
		$\beta_{10}$	9.94	0	1	2	3	4	5	6	7
		Estimated means	21.06	31.00	40.94	50.88	60.82	70.76	80.70	90.64	
	SEM (constrained)	$\beta_{00}$	21.06	1	1	1	1	1	1	1	1
		$\beta_{10}$	9.94	0	1	2	3	4	5	6	7
		Estimated means	21.06	31.00	40.94	50.88	60.82	70.76	80.70	90.64	
Piecewise growth model	HLM	$\beta_{00}$	21.4	1	1	1	1	1	1	1	1
		$\beta_{10}$	10.38	0	1	2	3	3	3	3	3
		Estimated means	21.76	31.55	41.34	51.13	60.92	70.71	80.5	90.29	
	SEM (constrained)	$\beta_{00}$	21.40	1	1	1	1	1	1	1	1
		$\beta_{10}$	10.38	0	1	2	3	3	3	3	3
		Estimated means	21.40	31.78	42.16	52.54	56.38	68.68	80.98	93.28	
Observed means	SEM (unconstrained)	$\beta_{00}$	21.20	1	1	1	1	1	1	1	1
		$\beta_{10}$	10.48	0	1	2	3	3	3	3	3
		Estimated means	21.08	32.91	42.38	52.65	56.09	69.18	81.7	92.7	
	SEM (unconstrained)	$\beta_{00}$	21.20	1	1	1	1	1	1	1	1
		$\beta_{10}$	10.48	0	1	2	3	3	3	3	3
		Estimated means	21.20	31.68	42.16	52.64	56.34	68.65	80.96	93.27	
	Observed means	21.08	32.91	42.38	52.65	56.09	69.18	81.7	92.7		

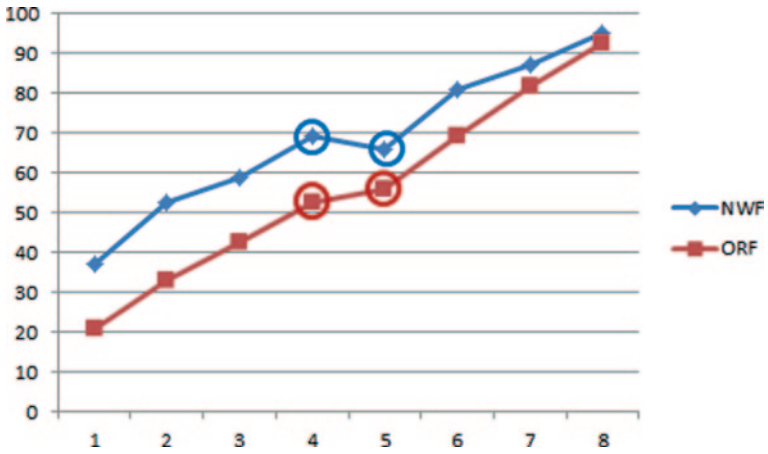


Fig. 10.5 The plot of means of the eight repeated measures on the ORF and NWF

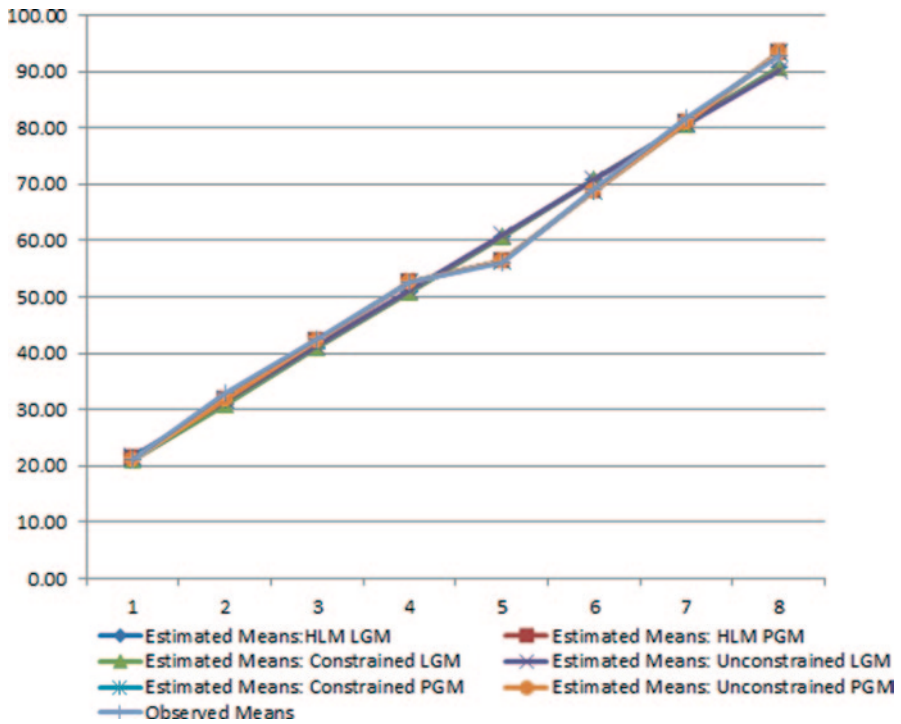


Fig. 10.6 Plots of observed means and estimated means at student level

It is essential to compare the model-predicted values to the sample values to ensure that the model is able to adequately recover the shape of the growth trajectory. Explicit comparisons of the model-predicted and observed values at each of the time points reveals the shortcomings in a given model.

Traditional multilevel approaches allow for the flexible treatment of time and provide elegant options for univariate growth models. But if we wish to consider two-dependent variables simultaneously within one model, it is generally easiest to employ structural equation modeling approaches to growth modeling. Our next set of models reintroduces NWF. Modeling NWF and ORF simultaneously helps us to better understand the linkages between the two variables. Traditional multilevel models are univariate. In contrast, the SEM approach accommodates multiple outcome variables seamlessly. Therefore, we fit the dual-process growth models using SEM.

### ***Model 2: The Unconditional Dual-Process Growth Curve Model***

In a dual-process growth model, we can model the concurrent growth on two outcome variables by fitting two growth trajectories simultaneously. So, in one sense, the dual-process model is akin to modeling side-by-side growth models. Yet the real advantage of the dual-process growth model is that it allows us to estimate the correlations among the growth parameters for two outcome variables (e.g., NWF and ORF). This allows us to answer many questions including: (1) how does the first-grade growth slope for NWF relate to students' ORF growth in first or second grades; and (2) how do NWF scores in the beginning of first grade relate to ORF growth in first grade, summer, or second grade?

Figure 10.5 plots observed means of the repeated measures on NWF and ORF. The plots of means indicate nonlinearities in the two trajectories. The slope changes appear to occur at the same time points. There are two elbows on the plots of both NWF and ORF, and therefore we introduce three separate slope pieces for each: the first slope represents the growth rate from time 1 to 4; the second slope represents the growth rate from time 4 to 5; and the third slope represents the growth rate from time 5 to 8. The plots suggest piecewise growth curve models might fit the data for both fluency constructs. However, the ORF trajectory looks much "cleaner" than the NWF trajectory. Whereas the ORF has three very distinct linear pieces, the NWF has some jaggedness, even within the three separate pieces.

To fit the dual process growth model, we estimate two separate latent variable growth models. Thus, there are eight latent variables in this dual-process piecewise growth curve model: two latent intercepts and six latent growth slopes. We allow the latent growth parameters (the intercepts and the slopes) to correlate with each other within and across each of the outcome variables.

Figure 10.7 presents the diagram of the measurement model, which is built to estimate the fully-crossed correlation matrix among the eight latent variables. In the measurement model, all the latent variables are allowed to correlate with each other. In this model, we had to constrain the correlation between the second ORF slope

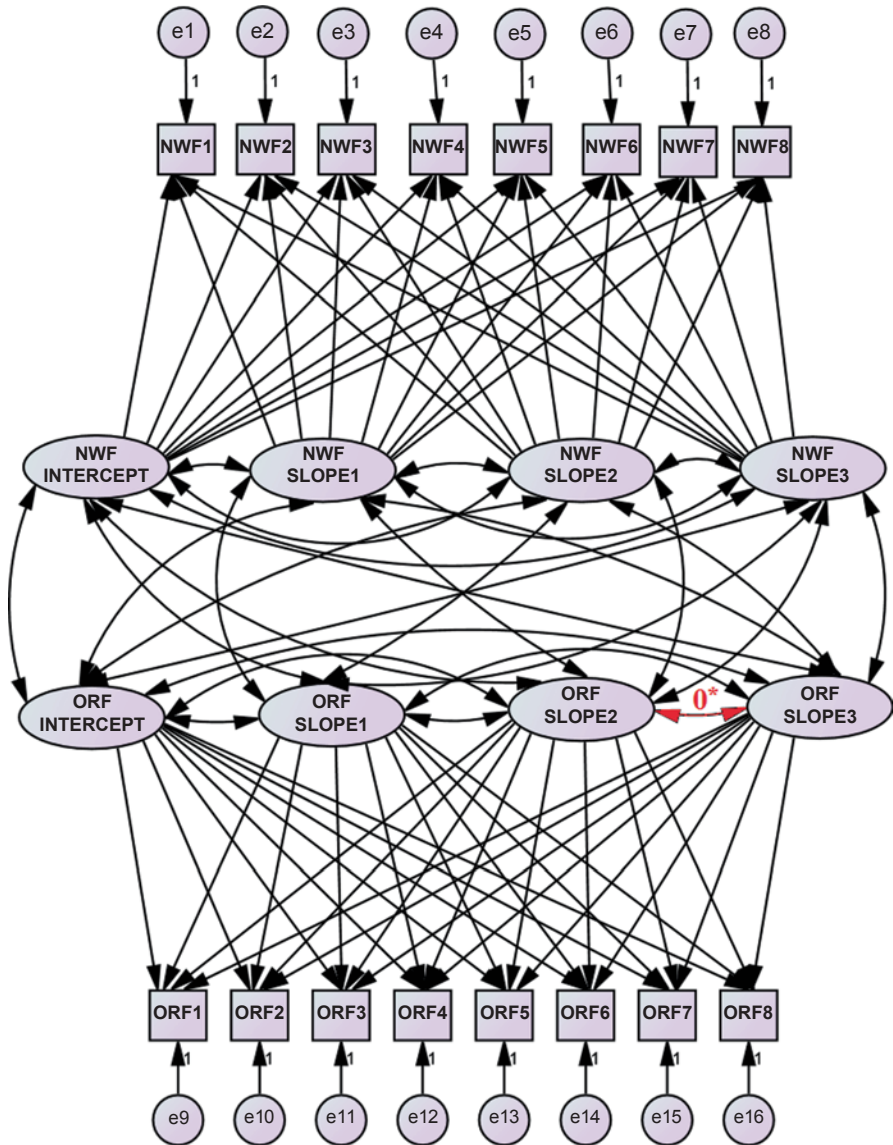


Fig. 10.7 Diagram of the measurement model

and the third ORF slope at zero to solve a “Heywood” case. Heywood cases are negative estimates of variances or correlation estimates greater than one (Kolenikov & Bollen, 2008). The covariance between the second and third NWF slope was also not statistically significantly different from 0.

Table 10.7 contains the means and covariances for the latent means and intercepts. The second (summer) growth slopes for both NWF and ORF are negatively related to the intercepts and the two school-year growth slopes.

**Table 10.7** Estimated means and covariance matrix for the latent variables

Mean								
	INW	SNW1	SNW2	SNW3	IOR	SOR1	SOR2	SOR3
	37.92	10.54	-2.33	9.81	21.23	10.47	3.66	12.33
Covariance matrix								
	INW	SNW1	SNW2	SNW3	IOR	SOR1	SOR2	SOR3
INW	380.57							
SNW1	25.17	32.99						
SNW2	-9.21	-10.16	104.08					
SNW3	17.67	9.88	2.16	40.97				
IOR	330.15	16.48	-8.23	16.90	384.62			
SOR1	54.27	22.79	-6.39	11.41	50.90	34.80		
SOR2	-36.28	-12.00	64.78	-10.66	-50.04	-5.57	65.43	
SOR3	3.32	5.55	-8.40	16.87	-9.02	6.89	0*	20.13

*INW* intercept for NWF, *SNW1* first grade slope for NWF, *SNW2* summer slope for NWF, *SNW3* second grade slope for NWF, *IOR* intercept for ORF, *SOR1* first grade slope for ORF, *SOR2* summer slope for ORF, *SOR3* second grade slope for ORF

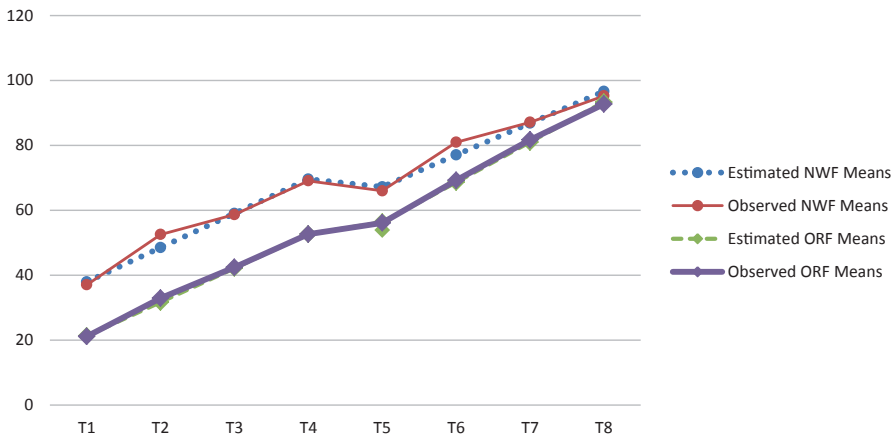
The covariance between *SOR3* and *SOR2* is constrained at zero

Table 10.8 contains the model predicted and sample observed means at each of the eight time points for NWF and ORF. The growth rates for NWF and ORF are quite similar in first grade, approximately 10.5 points ( $B_{SNW1} = 10.54$ ) and ( $B_{SOR1} = 10.47$ ). During the summer break, NWF decreases slightly ( $B_{SNW2} = -2.33$ ) while ORF increases slightly ( $B_{SOR2} = 3.66$ ). During second grade, the two slopes diverge slightly: the slope for NWF is slightly less in second grade than it was in first grade ( $B_{SNW3} = 9.81$ ) while ORF increases slightly ( $B_{SOR3} = 12.33$ ). So, in second grade, the slope of NWF is lower than it was in first grade, but the slope of ORF is higher than it was in first grade. Figure 10.8 compares the model predicted and sample observed means for ORF and NWF using the piecewise simultaneous growth model. As would be expected from Fig. 10.5, the three-piece model does a far better job reproducing the ORF data than it does with the NWF data.

Table 10.9 contains the bivariate correlations among the latent slopes and intercepts for ORF and NWF. Most of the correlations are statistically significant except the correlation between the summer and second-grade growth slopes of the NWF and the correlation between the summer and second-grade growth slopes of the ORF. The correlations between initial NWF and ORF is quite high ( $r = .86$ ) as are the correlations among the latent slopes measured during the same time periods. For example, the correlation between the first-grade ORF growth slope and first-grade NWF growth slope is .67; the correlation between the two summer reading slopes is .79, and the correlation between the two second-grade growth slopes is .59. Therefore, there is clearly a strong relation between students' growth on the two assessments.

**Table 10.8** NWF and ORF

Latent factors	Means	T1	T2	T3	T4	T5	T6	T7	T8
<b>NWF</b>									
INW	37.92	1	1	1	1	1	1	1	1
SNW1	10.54	0	1	2	3	3	3	3	3
SNW2	-2.33	0	0	0	0	1	1	1	1
SNW3	9.81	0	0	0	0	0	1	2	3
Estimated means	37.92	48.46	59.00	69.54	67.21	77.02	86.83	96.64	
Observed means	37.02	52.51	58.65	69.06	65.95	80.92	87.09	95.2	
Residual variance	87.44	170.24	173.67	188.47	180.03	324.62	265.84	337.61	
R-square		.813	.732	.779	.815	.826	.754	.82	.818
<b>ORF</b>									
IOR	21.23	1	1	1	1	1	1	1	1
SOR1	10.47	0	1	2	3	3	3	3	3
SOR2	3.66	0	0	0	0	1	1	1	1
SOR3	12.33	0	0	0	0	0	1	2	3
Estimated means	21.24	31.7	42.17	52.64	56.3	68.63	80.96	93.29	
Observed means	21.08	32.91	42.38	52.65	56.09	69.18	81.7	92.7	
Residual variance	26.04	45.91	72.97	57.61	56.22	65.63	89.48	77.52	
R-square		.937	.919	.909	.946	.943	.937	.922	.939



**Fig. 10.8** Plots of observed and estimated NWF means using the piecewise simultaneous growth model

***Model 3: Using the Growth Factors in a Regression Model***

The repeated measures follow a particular time sequence; therefore, using earlier latent growth factors from one construct to predict subsequent latent growth factors from another construct often makes conceptual sense. Fitting regression

**Table 10.9** Estimated correlation matrix for the latent intercepts and slopes

	INW	SNW1	SNW2	SNW3	IOR	SOR1	SOR2	SOR3
INW	1							
SNW1	.23	1						
SNW2	-.05	-.17	1					
SNW3	.14	.27	.03*	1				
IOR	.86	.15	-.04	.14	1			
SOR1	.47	.67	-.11	.30	.44	1		
SOR2	-.23	-.26	.79	-.21	-.32	-.12	1	
SOR3	.04	.22	-.18	.59	-.10	.26	0 <sup>a</sup>	1

Note: a -the covariance between SOR3 and SOR2 is constrained at zero; \* - Nonsignificant correlation.

paths between one latent factor and any other latent factors that precede it allow us to explore those explanatory relations. To demonstrate this technique, we recast the bivariate growth model as a simultaneous growth model in which earlier latent intercepts and growth slopes predicted later growth slopes. After deleting all the nonstatistically significant paths, we obtained the final model, shown in Fig. 10.9.

As we saw in the bivariate growth model, the correlation between the two latent intercepts was quite high ( $r = .86$ ), which indicates a strong relation between students' NWF scores and their ORF scores at the beginning of first grade. After controlling for the initial oral reading scores, initial NWF did positively predict both the first and second grade reading slopes. But after controlling for NWF, initial ORF negatively predicted first-grade NWF growth and did not predict second-grade NWF growth. This pattern of results suggests that initial NWF scores do have some value in predicting ORF. The paths between initial ORF and the three growth slopes are .093,  $-.349$ , and  $-.655$ , respectively. After controlling for initial NWF, and after controlling for the first-grade NWF and ORF growth slopes in the case of the second-grade ORF slope, the relation between initial ORF and the three growth slopes becomes increasingly negative. Although this may seem counter intuitive, it actually makes sense. Those who start out with the highest scores will not experience as much growth as those with lower scores during the summer and the second-grade school year, probably because growth on reading fluency slows after students reach a certain level of reading fluency. There are moderately high correlations (ranging from .5 to .6) between the disturbances for the three pairs of latent slopes. These correlations among the disturbances indicate that the unexplained variance in the ORF growth slope and the unexplained variance in the NWF growth slope are fairly strongly associated with each other.

According to Table 10.9, the growth slope of NWF in second grade (SNW3) correlates with all the other latent factors except the summer NWF growth slope (SNW2). But only the first-grade growth slope of NWF (SNW1) and the first-grade growth slope of ORF (SOR1) predict NWF growth in second grade. In other words, students' growth rates in both NWF and ORF during first grade help to predict their second-grade NWF growth; however, both of these coefficients are fairly small.



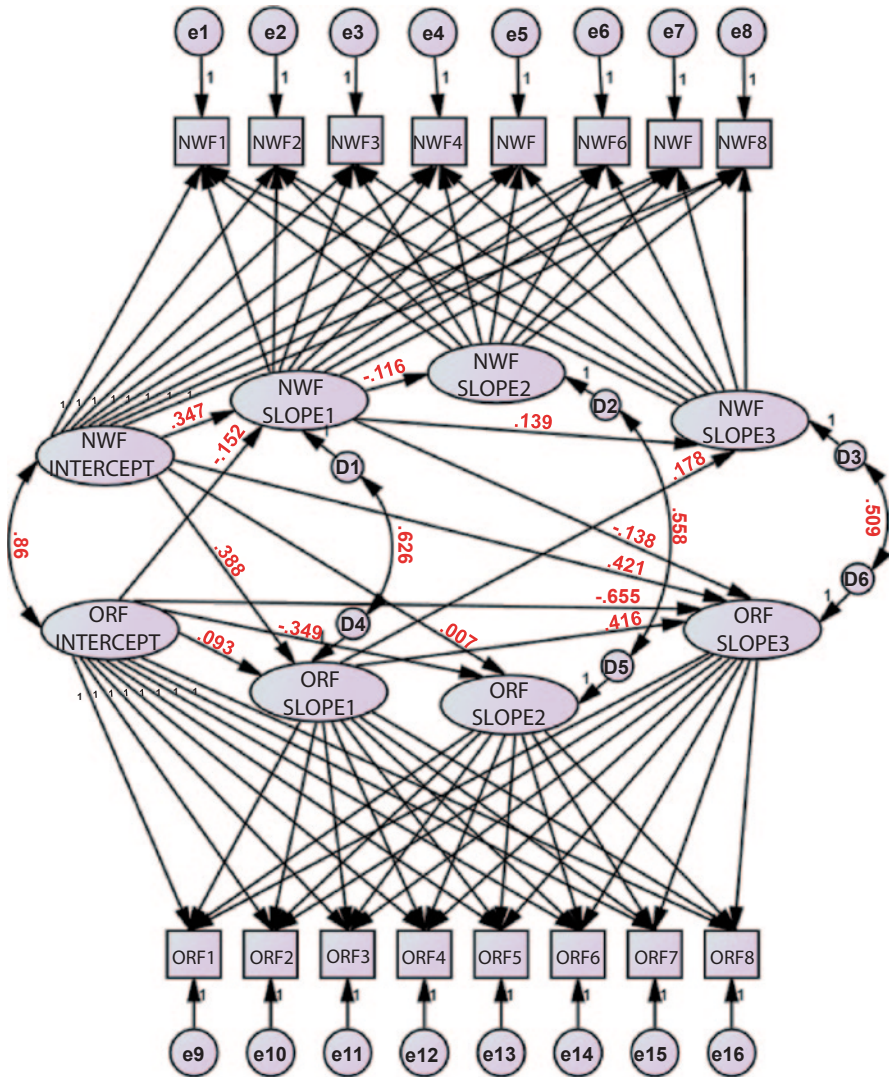


Fig. 10.9 Diagram of the final dual-process regression model

Moreover, after controlling for the growth rate in NWF and ORF during first grade, students' NWF growth rate during second grade is not correlated with students' initial NWF or ORF level (at the beginning of the first grade) or their growth rate during the summer vacation between first and second grades. In contrast, both initial ORF and NWF scores and first-grade reading slopes help to predict second-grade ORF slope.

Moreover, we also notice several negative path coefficients. For instance, the path coefficient between first-grade NWF growth slope and second-grade NWF growth slope is  $-.116$ . But the mean of the first growth slope of NWF is positive

and the second growth slope of NWF is negative. Therefore, the negative path coefficient indicates that students who make faster NWF growth during first grade regress slightly more on NWF during the summer vacation between first and second grades, but this is a very small effect. Further, neither the summer NWF growth slope nor the summer ORF slope is a good predictor of second-grade growth in either ORF or NWF. In summary, NWF does appear to provide some additional information to help us to predict students' ORF growth in second grade.

#### ***Model 4: Unconditional, Single-Process, Multilevel Growth Curve Model***

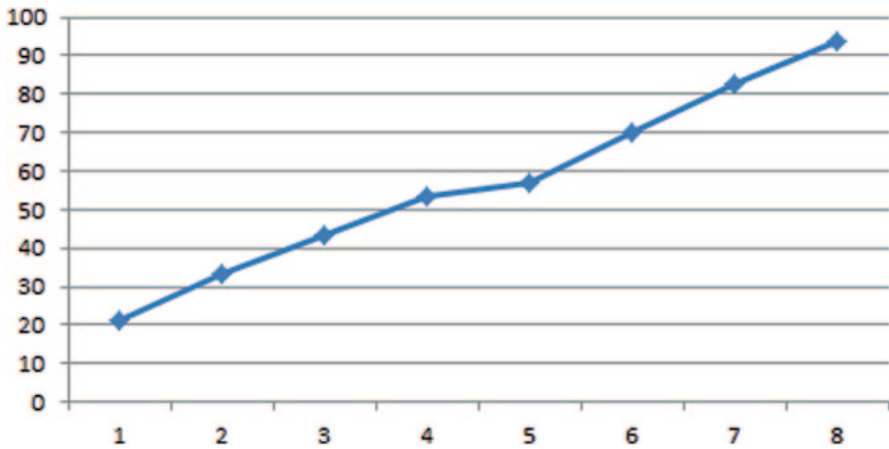
Growth models can also be extended to incorporate the clustering of students within schools. Such analyses can be conducted using two-level SEM models or three-level multilevel models. In this section, for simplicity, we illustrate multilevel growth modeling using a univariate outcome, ORF, again measured across eight time points. Students are clustered within schools, and we anticipate that students who attend the same school may be more similar to each other than students who attend different schools. Thus, estimates of the variance components and their standard errors for the fixed effects and the variance components should be more accurate when we take the clustered nature of the data into account. In addition, correlations between ORF scores across time may be different at the school level and the school level as students' ORF level and growth are conceptually different from schools' (aggregated) ORF and growth.

We examine the cluster effect of schools on students' ORF scores. Table 10.10 presents descriptive statistics for ORF as well as correlations among the eight repeated measures on ORF at the student level and at the school level. First, the correlations among the repeated measures are moderately high (ranging from .64 to .93 within-school and from .65 to .97 between-schools). Clearly, ORF measures over time are strongly related both within- and between-schools. Second, the within-school correlations tend to be lower than the corresponding between-school correlations. Third, the correlations among scores decrease as the time intervals between observations increase. Fourth, as we have seen previously, the ORF means increase over time and the growth rates are not consistent over time, especially during the summer vacation between first and second grade. According to the plot in Fig. 10.10, schools' mean ORF changes very little during the summer vacation period. Finally, the intraclass correlations for ORF at each time point range from .06 to .08, which indicates that 6–8% of the variance in ORF lies across schools. In other words, school can explain 6–8% of the variance in ORF at any given time point; 92–94% of the ORF variance lies within-schools. Although these are fairly modest ICCs, the best approach to accounting for the dependence in the data is to use multilevel growth modeling.

Based on the piecewise growth curve model, we also present two approaches to analyze the multilevel growth curve model: the multilevel modeling approach and the SEM approach. Table 10.11 presents these two approaches side by side

**Table 10.10** Descriptive statistics and correlations among the repeated measures of ORF at within- and between-school levels

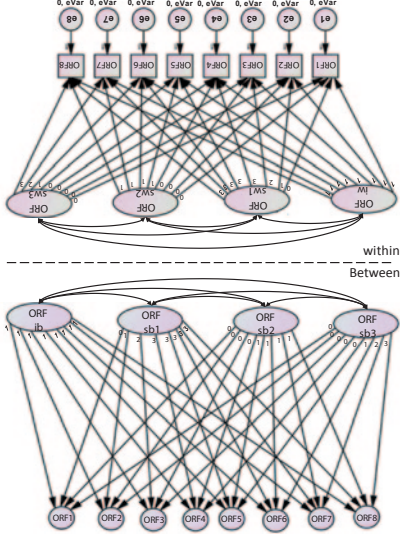
	ORF-1	ORF-2	ORF-3	ORF-4	ORF-5	ORF-6	ORF-7	ORF-8
Within-school								
ORF-1	1.00							
ORF-2	.91	1.00						
ORF-3	.85	.93	1.00					
ORF-4	.80	.88	.93	1.00				
ORF-5	.76	.85	.89	.92	1.00			
ORF-6	.72	.81	.86	.90	.92	1.00		
ORF-7	.66	.76	.82	.86	.89	.93	1.00	
ORF-8	.64	.73	.79	.84	.87	.91	.92	1.00
Variance	360.85	601.97	882.15	863.95	931.98	928.10	1088.32	1135.38
Between-school								
ORF-1	1.00							
ORF-2	.95	1.00						
ORF-3	.91	.97	1.00					
ORF-4	.82	.92	.96	1.00				
ORF-5	.82	.89	.93	.94	1.00			
ORF-6	.77	.85	.90	.92	.95	1.00		
ORF-7	.71	.79	.86	.88	.91	.96	1.00	
ORF-8	.65	.75	.82	.84	.87	.93	.96	1.00
Means	21.35	33.40	43.35	53.34	56.76	69.99	82.51	93.62
Variance	25.45	49.18	76.21	74.46	67.49	70.95	83.93	77.82
ICC	.066	.076	.080	.079	.068	.071	.072	.064



**Fig. 10.10** Plot of school mean ORF scores

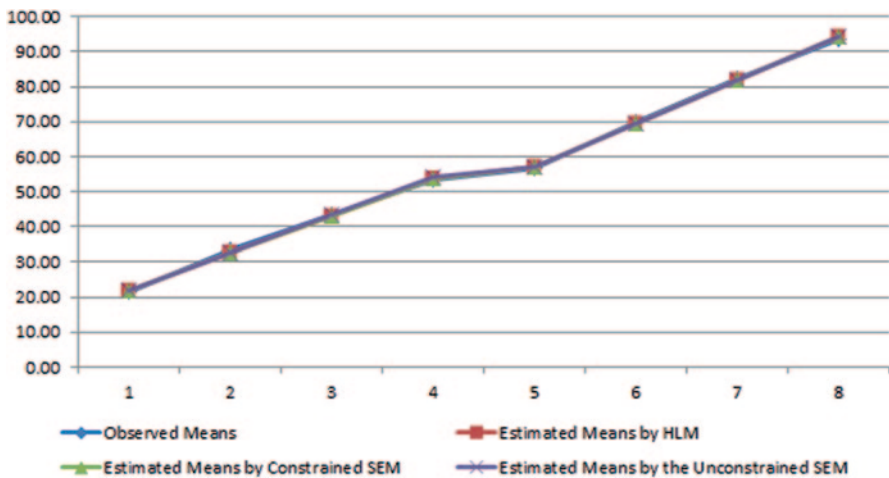
for comparison. Figure 10.11 plots both the observed means and estimated school means. In the corresponding SEM model, all the residual variances are constrained to be equal across all time points within students and constrained to be zero across

**Table 10.11** Two approaches to modeling growth when people are nested within clusters

Three-level univariate HLM model	Two-level multivariate SEM model
<p>Level-1 model:</p> $y_{ijk} = i_{0jk} + s_{1jk} * (x_1)_{ijk} + s_{2jk} * (x_2)_{ijk} + s_{3jk} * (x_3)_{ijk} + e_{ijk}$ <p>Level-2 Model:</p> $i_{0jk} = \beta_{00k} + r_{0jk}$ $s_{1jk} = \beta_{10k} + r_{1jk}$ $s_{2jk} = \beta_{20k} + r_{2jk}$ $s_{3jk} = \beta_{30k} + r_{3jk}$ <p>Level-3 Model:</p> $\beta_{00k} = \gamma_{000} + u_{00k}$ $\beta_{10k} = \gamma_{100} + u_{10k}$ $\beta_{20k} = \gamma_{200} + u_{20k}$ $\beta_{30k} = \gamma_{300} + u_{30k}$	
<p>Four fixed effects:</p> <ul style="list-style-type: none"> <li><math>\gamma_{000}</math> : the expected initial score</li> <li><math>\gamma_{100}</math> : the expected growth rate from time 1 to time 4</li> <li><math>\gamma_{200}</math> : the expected growth rate from time 4 to time 5</li> <li><math>\gamma_{300}</math> : the expected growth rate from time 5 to time 8</li> </ul> <p>Nine random effects:</p> <p>Level-1, within students:</p> <p>Variance of <math>e_{ijk}</math> (<math>\sigma^2</math>): the variance within students across time, which is assumed to be homogenous</p> <p>Level-2, between students within schools: Variance of <math>r_{0jk}</math> (<math>\tau(\pi)_{00}</math>): the variance in the initial ORF scores crossing students within schools</p> <p>Variance of <math>r_{1jk}</math> (<math>\tau(\pi)_{11}</math>): the variance in the first slope between students and within schools, which represents the growth rate from time 1 to time 4</p> <p>Variance of <math>r_{2jk}</math> (<math>\tau(\pi)_{22}</math>): the between-student variance in the second slope within schools, which represents the growth rate from time 4 to time 5</p> <p>Variance of <math>r_{3jk}</math> (<math>\tau(\pi)_{33}</math>): the between-student variance within schools in the third slope, which represents the growth rate from time 5 to time 8</p> <p>Level-3, Between schools:</p> <p>Variance of <math>u_{00k}</math> (<math>\tau(\beta)_{00}</math>): the between-schools variance in the initial ORF scores</p>	<p>“Within” and “Between” indicate within and between the “CLUSTER”. In this two-level SEM model, the “CLUSTER” is “SCHOOL”</p> <p>The level-1 residual variance within students is constrained to be equal across time points and level-2 residual variance between schools is constrained to zero, which are the two default assumptions in the HLM program</p> <p>To construct a SEM model identical to the HLM model, these two constraints are introduced</p> <p>The residual variance of the eight observed variables (or dependent variables) within students corresponds to the level-1 error variance (<math>\sigma^2</math>) in the HLM model</p> <p>The means of the latent intercept or the latent slopes are ONLY estimated at the highest level, which is the “%BETWEEN%” level in this case. Therefore, the mean of “ib” is corresponding to the fixed effect of <math>\beta_{00k}</math> (<math>\gamma_{000}</math>) in the HLM model, the mean of “sb1”, “sb2”, and “sb3,” respectively correspond to the fixed effects of <math>\beta_{10k}</math> (<math>\gamma_{100}</math>), <math>\beta_{20k}</math> (<math>\gamma_{200}</math>), and <math>\beta_{30k}</math> (<math>\gamma_{300}</math>)</p> <p>The variance of “sw1”, “sw2”, and “sw3” correspond to the level-2 random effects of <math>s_{1jk}</math> (<math>\tau(\pi)_{11}</math>), <math>s_{2jk}</math> (<math>\tau(\pi)_{22}</math>), and <math>s_{3jk}</math> (<math>\tau(\pi)_{33}</math>) in the HLM model. The variance of “sb1”, “sb2”, and “sb3” correspond to the level-3 variance components of <math>u_{10k}</math> (<math>\tau(\beta)_{11}</math>), <math>u_{20k}</math> (<math>\tau(\beta)_{22}</math>), and <math>u_{30k}</math> (<math>\tau(\beta)_{33}</math>) in the HLM model</p>

**Table 10.11** (continued)

Three-level univariate HLM model	Two-level multivariate SEM model
Variance of $u_{10k}(\tau(\beta)_{11})$ : the between-schools variance in the first slope, which represents the growth rate from time 1 to time 4 Variance of $u_{20k}(\tau(\beta)_{22})$ : the between-schools variance in the second slope, which represents the growth rate from time 4 to time 5 Variance of $u_{30k}(\tau(\beta)_{33})$ : the between-schools variance in the third slope, which represents the growth rate from time 5 to time 8	



**Fig. 10.11** Plots of observed and estimated school mean ORF scores

schools. After considering this standard approach to multilevel growth modeling, we present an SEM model without these constraints and examine the influence of the constraints on the estimates of fixed and random effects. Therefore, we present three multilevel growth models: a three-level HLM model, a constrained two-level SEM model identical to the HLM model, and an unconstrained two-level SEM model.

Table 10.12 summarizes the estimated variance components and Table 10.13 summarizes the estimated fixed effects obtained from the three models. The unconstrained model suggests that residual variances do differ across the eight time points. Comparing the estimated variance components from the multilevel growth model (Table 10.12) and the ones in Table 10.5, the estimated residual variances ( $\sigma^2$ ) are roughly the same from the constrained models with or without consideration of the cluster effect of schools. The between-student variance components in the two-level HLM model (single-level SEM model) are larger than the corresponding variance components in the three-level HLM model (or the two-level

**Table 10.12** Random effects

HLM		Constrained SEM		Unconstrained SEM	
Within- student					
$\delta^2$	65.20	Residual variance constrained as equal	64.88	Residual variance ORF1	24.79
				Residual variance ORF2	43.69
				Residual variance ORF3	69.89
				Residual variance ORF4	54.88
				Residual variance ORF5	66.42
				Residual variance ORF6	60.59
				Residual variance ORF7	84.22
				Residual variance ORF8	80.09
Between-student within-school level					
$\tau(\pi)_{00}$	362.45	Variance of i	372.02	Variance of i	369.77
$\tau(\pi)_{11}$	25.85	Variance of s1	25.49	Variance of s1	31.49
$\tau(\pi)_{22}$	40.64	Variance of s2	39.15	Variance of s2	45.31
$\tau(\pi)_{33}$	17.05	Variance of s3	16.26	Variance of s3	14.64
Between-school level					
$\tau(\beta)_{00}$	31.45	Variance of i	31.27	Variance of i	26.84
$\tau(\beta)_{11}$	3.22	Variance of s1	3.25	Variance of s1	3.27
$\tau(\beta)_{22}$	8.87	Variance of s2	8.96	Variance of s2	6.15
$\tau(\beta)_{33}$	2.09	Variance of s3	2.13	Variance of s3	1.40

SEM model). This is because the three-level HLM model (two-level SEM model) partitions the between-student variance into two components, one of which lies between students within schools and the other of which lies between schools. The between-school variance components are relatively small as compared to the corresponding variance components between students within schools in this case. But including the cluster effect of schools should provide more accurate estimates of the standard error. Table 10.13 provides the parameter estimates for the three multilevel growth models. The parameter estimates for the multilevel growth models are quite similar to the parameter estimates for the piecewise growth models that we presented earlier.

### *Model Fit Comparisons of the Growth Models*

In mean and covariance structure (MACS) SEM, the number of observed variables and the number of levels in the model determines the maximum number of parameters that can be estimated. For single-level models, the maximum number of parameters that can be estimated is  $\frac{n \times (n+1)}{2} + n$  or  $\frac{n \times (n+3)}{2}$ . For multilevel models, the maximum number of parameters is  $n + l \times \frac{n \times (n+1)}{2}$ , where n is the number of observed variables, l is the number of levels. Table 10.14 demonstrates

**Table 10.13** Observed and estimated means at each time point

Approach	Model	Fixed effects	Mean	T1	T2	T3	T4	T5	T6	T7	T8
HLM	Piecewise growth model	$\beta_{00}$	21.95	1	1	1	1	1	1	1	1
		$\beta_{10}$	10.59	0	1	2	3	3	3	3	3
		$\beta_{20}$	3.52	0	0	0	0	1	1	1	1
		$\beta_{30}$	12.31	0	0	0	0	0	1	2	3
		Estimated means		21.95	32.54	43.13	53.72	57.24	69.55	81.86	94.17
SEM	Constrained piecewise growth model	$\beta_{00}$	21.96	1	1	1	1	1	1	1	1
		$\beta_{10}$	10.59	0	1	2	3	3	3	3	3
		$\beta_{20}$	3.51	0	0	0	0	1	1	1	1
		$\beta_{30}$	12.31	0	0	0	0	0	1	2	3
		Estimated means		21.96	32.55	43.14	53.73	57.24	69.55	81.86	94.17
	Unconstrained piecewise growth model	$\beta_{00}$	21.71	1	1	1	1	1	1	1	1
		$\beta_{10}$	10.79	0	1	2	3	3	3	3	3
		$\beta_{20}$	3.17	0	0	0	0	1	1	1	1
		$\beta_{30}$	12.35	0	0	0	0	0	1	2	3
		Estimated means		21.71	32.5	43.29	54.08	57.25	69.6	81.95	94.30
Observed means of the school-means											
Observed grand means											
				21.35	33.4	43.35	53.34	56.76	69.99	82.51	93.62
				21.08	32.91	42.38	52.65	56.09	69.18	81.7	92.70



**Table 10.14** Calculation of degrees of freedom

Observed variable	Model	Estimated variables	df
ORF	Single-process linear growth model (latent factors: i s1)		
	HLM model constrained SEM model	$n = 8, l = 1$ Max number: $8(8 + 3)/2 = 44$ Number of error variance ( $\sigma^2$ ): 1 Number of variances in latent factors: 2 Number of means: 2 Number of covariances among latent factors: 1	$44 - 1 - 2 - 2 - 1 = 38$
	Unconstrained SEM model	$n = 8, l = 1$ Max number: $8(8 + 3)/2 = 44$ Number of error variance: 8 Number of variances in latent factors: 2 Number of means: 2 Number of covariances among latent factors: 1	$44 - 8 - 2 - 2 - 1 = 31$
	Single-process piecewise growth model (latent factors: i s1 s2 s3)		
	HLM Model constrained SEM model	$n = 8, l = 1$ Max number: $8(8 + 3)/2 = 44$ Number of error variance ( $\sigma^2$ ): 1 Number of variances in latent factors: 4 Number of means: 4 Number of covariances among latent factors: 6	$44 - 1 - 4 - 4 - 6 = 29$
	Unconstrained SEM model	$n = 8, l = 1$ Max number: $8(8 + 3)/2 = 44$ Number of error variance ( $\sigma^2$ ): 8 Number of variances in latent factors: 4 Number of means: 4 Number of covariances among latent factors: 6	$44 - 8 - 4 - 4 - 6 = 22$
ORF and NWF	Dual-process piecewise growth model (latent factors: iNW sNW1 sNW2 sNW3 iOR sOR1 sOR2 sOR3)		
	Unconstrained SEM model (there is one constraint made to solve the Heywood case)	$n = 16, l = 1$ Max number: $16(16 + 3)/2 = 152$ Number of error variances ( $\sigma^2$ ): 16 Number of variances for latent factors: 8 Number of means: 8 Number of covariances among latent factors: $(28 - 1)$	$152 - 16 - 8 - 8 - 27 = 93$

**Table 10.14** (continued)

Observed variable	Model	Estimated variables	df
ORF	Single-process multilevel piecewise growth model (latent factors: i s1 s2 s3)		
	HLM model constrained SEM model	$n = 8, l = 2$ Max number: $8 + 2 \times 8(8 + 1)/2 = 80$ Number of error variances ( $\sigma^2$ ): 1 Number of variances for Latent Factors: 8 Number of means: 4 Number of covariances among latent factors: 12	$80 - 1 - 8 - 4 - 12 = 55$
	Constrained SEM model	$n = 8, l = 2$ Max number: $8 + 2 \times 8(8 + 1)/2 = 80$ Number of error variance ( $\sigma^2$ ): 16 Number of variances in latent factors: 8 Number of means: 4 Number of covariances among latent factors: 12	$80 - 16 - 8 - 4 - 12 = 40$

how to calculate the degrees of freedom for the different models that we have presented.

If two models are nested, we can use the model chi-square or the model deviance to compare the two models directly. Two models are nested when one model is a subset of the other (Kline, 1998). The simpler model always has a higher chi-square or deviance than the more complex model. In large samples, the difference between the deviances of two hierarchically nested models is distributed as an approximate chi-square distribution with degrees of freedom equal to the difference in the number of parameters being estimated between the two models (de Leeuw, 2004). Subtracting the chi-square or deviance of the simpler model from the chi-square or deviance from the more complex model provides a change score that can be compared to the critical value of chi-square with degrees of freedom equal to the difference in the degrees of freedom between the two models. In evaluating model fit using the chi-square difference test, the more parsimonious model is preferred, as long as it does not result in significantly worse fit. In other words, if the model with the larger number of parameters (fewer degrees of freedom) fails to reduce the chi-square or deviance by a substantial amount, the more parsimonious model is retained. Therefore, when the change in deviance exceeds the critical value of chi-square with degrees of freedom equal to the difference in the number of parameters being estimated between the two models, we favor the more complex model. But if the more complex model does not result in a statistically significant reduction in the deviance statistic, we favor the more parsimonious model.

**Table 10.15** Chi-square test on nested models (from the most parsimonious to the most complex model)

Model	$\chi^2$	df	$\Delta\chi^2$	$\Delta$ df	$\Delta\chi^2$	$\Delta$ df	P-value
Constrained linear growth model	31671.80	38					
Unconstrained linear growth model	27205.10	31	4466.7	7			<.001
Constrained piecewise growth model	10394.25	29	21277.55	9			<.001
Unconstrained piecewise growth model	7405.53	22	24266.27	7	2988.7	7	<.001

We conducted chi-square difference tests to compare the fit of the constrained linear growth model to the other three models. In addition, we conducted a chi-square difference test to compare the fit of the two piecewise growth models. The results are presented in Table 10.15. Not surprisingly, the results indicate that the piecewise growth models fit better than the linear growth models do. In addition, the chi-square comparison between the constrained and unconstrained piecewise growth models indicates that unconstrained model fits better, which suggests that the residual variances are not homogeneous across time. Table 10.16 summarizes the model fit indices of all our models.

## Summary

What have we learned about reading fluency using these approaches? First and foremost, growth in reading fluency does not follow a linear trajectory across first and second grades. Instead, growth is faster during the school year and slower during the summer. Also, the growth rates in first and second grades differ from each other. For ORF, the growth rate is higher in second grade than it is in first grade, whereas for NWF, the opposite is true. ORF and NWF growth rates are fairly strongly related to each other. We can do a better job predicting ORF growth rates than we can predicting NWF growth rates. School only explains a small amount of the between-student variability in growth rates. Finally, the within-person residuals do differ across time in our models, but constraining them to be equal across time has very little effect on our parameter estimates.

We hope that this introduction to growth modeling within the multilevel and SEM frameworks provides a starting point for fluency researchers who are interested in analyzing change or growth. Although we have provided several illustrations of growth models for fluency within this chapter, we have provided only a small sample of the possible growth models that can be fit to fluency data. In addition, other longitudinal models, such as autoregressive, cross-lagged, or latent change score models may prove even more beneficial in terms of understanding the nature of fluency growth and development across time.

**Table 10.16** Fit indices for model testing

Outcome	Growth curve model	CFI	TLI	BIC	RMSEA	$\chi^2$	df
ORF	Unconditional single-process single level linear growth model						
	HLM model constrained SEM model	.87	.91	1166288.23	.211	31671.80	38
	Unconstrained SEM model	.89	.90	1161890.37	.217	27205.10	31
	Unconditional single-process piecewise growth model						
	HLM model constrained SEM model	.96	.96	1145099.19	.138	10394.25	29
	Unconstrained SEM model	.97	.96	1142179.31	.134	7405.53	22
ORF & NWF	Dual-process						
	Unconstrained SEM model	.97	.96	2387179.06	.091	14402.33	93
	Regression model	.97	.96	2387393.87	.087	14715.49	103
Unconditional single-process multilevel							
ORF	HLM model constrained SEM model	.95	.95	927704.40	.082	5475.674	55
	Unconstrained SEM model	.96	.95	923943.45	.083	4109.138	40

Note: *TLI* Tucker-Lewis index, *RMSEA* root mean square error of approximation

## Appendix A: Computing the Estimated Means at Each Time Point

Using the estimated fixed effects from the HLM piecewise growth model, we demonstrate how to calculate the estimated means at each time point. In general, for student  $i$ ,  $Y_i$  is a  $1 \times 8$  vector representing the eight observations across time,  $B_i$  is a  $1 \times 4$  vector representing the four estimated fixed effects from the HLM piecewise growth model, and  $X_i$  is a  $4 \times 8$  vector representing the factor loadings of the latent intercept and the three latent growth slopes. Then, the calculation is:

$$E[Y_i] = E[B_i X_i + e_i] = \frac{n!}{r!(n-r)!} = (21.40 \quad 10.38 \quad 3.84 \quad 12.30) \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 3 & 3 & 3 & 3 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 \end{pmatrix}$$

$$= \begin{pmatrix} 21.40 \times 1 + 10.38 \times 0 + 3.84 \times 0 + 12.30 \times 0 \\ 21.40 \times 1 + 10.38 \times 1 + 3.84 \times 0 + 12.30 \times 0 \\ 21.40 \times 1 + 10.38 \times 2 + 3.84 \times 0 + 12.30 \times 0 \\ 21.40 \times 1 + 10.38 \times 3 + 3.84 \times 0 + 12.30 \times 0 \\ 21.40 \times 1 + 10.38 \times 3 + 3.84 \times 1 + 12.30 \times 0 \\ 21.40 \times 1 + 10.38 \times 3 + 3.84 \times 1 + 12.30 \times 1 \\ 21.40 \times 1 + 10.38 \times 3 + 3.84 \times 1 + 12.30 \times 2 \\ 21.40 \times 1 + 10.38 \times 3 + 3.84 \times 1 + 12.30 \times 3 \end{pmatrix}$$

$$= (21.40 \quad 31.78 \quad 42.16 \quad 52.54 \quad 56.38 \quad 68.68 \quad 80.98 \quad 93.28).$$

Taking time 6 as an example, the estimated mean ORF is equal to:  
 $21.40 \times 1 + 10.38 \times 3 + 3.84 \times 1 + 12.30 \times 1 = 68.68.$

## References

Bast, J., & Reitsma, P. (1997). Matthew effects in reading: A comparison of latent growth curve models and simplex models with structured means. *Multivariate Behavioral Research, 32*, 135–167.

Bast, J., & Reitsma, P. (1998). Analyzing the development of individual differences in terms of Matthew effects in reading: Results from a Dutch longitudinal study. *Developmental Psychology, 34*, 1373–1399.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken: Wiley Interscience.

Campbell, D. T., & Kenny, D. A. (1999). *A primer of regression artifacts*. New York: Guilford Press.

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*(2), 163–176. doi:10.1177/0022219408326219.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology, 57*, 505–528.

Crowe, E., Connor, C., & Petscher, Y. (2009). Examining the core: Relations among reading curricula, poverty, and first through third grade reading achievement. *Journal of School Psychology, 47*(3), 187–214. doi:10.1016/j.jsp.2009.02.002.

Curran, P. J., Lee, T. H., Howard, A. L., Lane, S. T., & MacCallum, R. C. (2012). Disaggregating within-person and between-person effects in multilevel and structural equation growth models. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 217–253). Charlotte: Information Age Publishing.

Duncan, T. E., Duncan, S. C., & Stryker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications (Second edition)*. Mahwah: Lawrence Erlbaum.

Frees, E. W. (2004). *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press.

- Gagne, R. M., Wager, W. W., Golas, K. C., Keller, J. M., & Russell, J. D. (2005). *Principles of instructional design* (5th ed.). Belmont, CA: Thomson Learning Inc.
- Hsiao, C. (2003). Analysis of panel data, 2nd edn. Econometric society monographs, vol 34. Cambridge University Press, Cambridge.
- Kenny, D. (1974). A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin*, *82*, 342–362.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
- Kline, R. B., & ebrary, I. (2011). Principles and practice of structural equation modeling. New York: Guilford Press.
- Kolenikov, S., & Bollen, K. A. (2008). Testing negative error variances: Is a Heywood Case a symptom of misspecification? *University of North Carolina*.
- Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: Guilford Press.
- Logan, J. R., & Petscher, Y. (2010). School profiles of at-risk student concentration: Differential growth in oral reading fluency. *Journal of School Psychology*, *48*(2), 163–186. doi:10.1016/j.jsp.2009.12.002.
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual review of psychology*, *60*, 577–605.
- McCoach, D. B., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model of children's reading growth during the first 2 years of school. *Journal of Educational Psychology*, *98*(1), 14.
- McCoach, D. B., Madura, J. P., Rambo-Hernandez, K. E., O'Connell, A. A., & Welsh, M. E. (2013). Longitudinal data analysis. In *Handbook of quantitative methods for educational research* (pp. 199–230). SensePublishers.
- McNamara, J. K., Scissons, M., & Gutknecht, N. (2011). A longitudinal study of kindergarten children at risk for reading disabilities: The poor really are getting poorer. *Journal of Learning Disabilities*, *44*(5), 421–430. doi:10.1177/0022219411410040.
- Muthén, L. K., & Muthén, B. O. (1998). 2012. Mplus User's Guide, 7.
- Puranik, C. S., Petscher, Y., Al Otaiba, S., Catts, H. W., & Lonigan, C. J. (2008). Development of oral reading fluency in children with speech or language impairments: A growth curve analysis. *Journal of Learning Disabilities*, *41*(6), 545–560. doi:10.1177/0022219408317858.
- Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). London: Sage Publications.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- Skrondal, A., & Rabe-Hesketh, S. (2008). Multilevel and related models for longitudinal data. In J. deLeeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 275–300). New York: Springer.
- Speece, D. L., & Ritchey, K. D. (2005). A longitudinal study of the development of oral reading fluency in young children at risk for reading failure. *Journal of Learning Disabilities*, *38*(5), 387–399. doi:10.1177/00222194050380050201.
- Stoel & Garre (2011). Growth curve analysis using multilevel regression and structural equation modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 97–111). New York: Routledge.
- Thorndike, R. L. (1966). Intellectual status and intellectual growth. *Journal of Educational Psychology*, *57*(3), 121.
- Willett, J. B. (1989). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422.

# Chapter 11

## Introduction to Latent Class Analysis for Reading Fluency Research

Jessica A. R. Logan and Jill M. Pentimonti

The practice of splitting people, places, or things into groups is common in all aspects of our lives. Doctors group their patients to decide if someone is at risk for a particular disease. Psychologists group their patients into whether or not they can be diagnosed with a mental disorder. Policymakers make decisions about whether programs are failing or succeeding, and teachers have to decide whether or not a student needs intervention. While some of those decisions are more data based than others, the common and sometimes even subconscious division of constructs into groups is an integral part of many things we do. However, splitting people into groups can be problematic; a fact that becomes clear as soon as you begin to operationalize how to split people into those groups. Imagine a scenario where a teacher needs to determine which of the students in her first grade class are in of a reading intervention. The teacher could choose to use interim progress monitoring assessments such as the oral reading fluency (ORF) subtest of dynamic indicators of basic early literacy skills assessment (*DIBELS*; Good, Kaminski, Smith, Laimon, & Dill, 2001). In this test, a midyear first grader who reads 20 words correctly in 1 min is considered to be reading well, while those reading more slowly or inaccurately are considered to be at some risk for reading failure. This type of classification implies that students who read 19 words correctly per minute (WCPM) are qualitatively different than those who read 20 WCPM. To extend this to the teacher in our scenario, she should provide intervention for those who read 19 WCPM, but not for those who read 20 WCPM. Realistically, this would not be the only factor a teacher would consider. The DIBELS has several subtests, and each one has a suggested benchmark or criteria that indicate risk. A student who can correctly name fewer than 27 letters exactly in 1 minute is also considered to be at risk. Just like with ORF, this benchmark system indicates that a student reading fewer than 26 letters correctly

---

J. A. R. Logan (✉) · J. M. Pentimonti  
The Crane Center for Early Childhood Research and Policy, Ohio State University, Columbus,  
OH 43210, USA  
e-mail: logan.251@osu.edu

J. M. Pentimonti  
e-mail: JPentimonti@ehe.osu.edu

© Springer Science+Business Media, LLC 2016  
K. D. Cummings, Y. Petscher (eds.), *The Fluency Construct*,  
DOI 10.1007/978-1-4939-2803-3\_11



in 1 minute is in need of intervention, while a student reading 27 letters is not. It is easy to imagine a scenario where a child scores below the benchmark in one test but not in the other, in which case the decision of whether to spend resources intervening with this student is more complex.

The goal of this chapter is to describe latent class analysis (LCA), which is a technique that is in essence about splitting a sample of people, places, or things into meaningful groups. As mentioned above, there are clear instances when grouping people is necessary. Such is the case with the example teacher who has to decide whether to provide extra intervention for a child. LCA is an excellent method as it lets you identify groups based on the data you have rather than an arbitrary idea of what constitutes membership in a group.

## Conceptual Introduction

The goal of this chapter is to introduce the reader to the concepts underlying an LCA, specifically in terms of how it can be used with measures of reading fluency. In later sections, an application of LCA using reading fluency data is demonstrated. The topic will first be introduced here in the context of a construct that has been more commonly assessed with LCA in the research literature; the classification of a mental disorder. Consider the plight of the psychologist who must decide whether a person has attention deficit/hyperactivity disorder (ADHD) or not. In any given sample of people (let us use students as they are most likely to obtain a diagnosis of ADHD), we can assume that four different groups exist. Some students will have the inattentive type, some the hyperactivity type, some with the combined type, and last there will be some students who do not have ADHD. The way students are typically given a classification into one of these categories is to have a clinician determine how many of the nine indicators of hyperactivity and nine indicators of attention problems (included in the Diagnostic and Statistical Manual of Mental Disorders; DSM-V) apply to the student (American Psychiatric Association, 2013). If a student displays six or more behaviors (in the right categories) then they are diagnosed with the corresponding type of ADHD. The fact that six items is the cutoff, rather than five or seven, is a seemingly arbitrary distinction.

Instead of this arbitrary cutoff approach, several researchers have instead used LCA to divide students into groups based on their ADHD symptoms (e.g., Hudziak, Heath, Madden, Reich, Bucholz, Slutske, Bierut, Neuman, & Todd, 1998; Rasmussen, Neuman, Heath, Levy, Hay, & Todd, 2002). The LCA method allows researchers in this area to assign each person to a group based on their responses to the same 18 items mentioned earlier. Because the test was designed to assess two different constructs, an LCA should theoretically identify four groups based on each student's probability of endorsing each of the items: one group with a high probability of endorsing the items that measure hyperactivity, one with a high probability of endorsing the inattention items, one with a high probability on all items, and one with a low probability on all items.

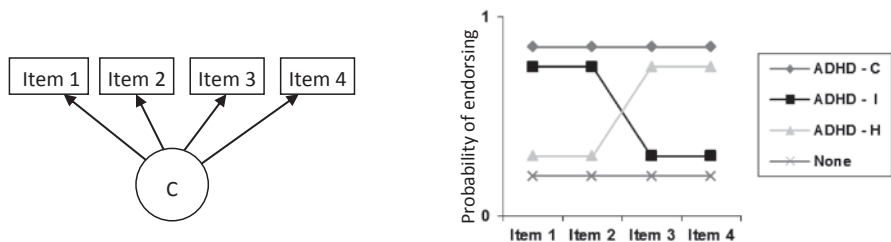


Fig. 11.1 Illustrated example for the analysis of ADHD symptoms

A simplified representation of the analysis is depicted graphically in Fig. 11.1, where the first two boxes on the left side of the figure represent items on the test that measure inattention, while the second two represent those items that measure hyperactivity. You will notice that the four squares representing the observed variables, all have arrows into them from a circle that is labeled “C.” Traditional structural-equation-modeling notation holds that observed variables are represented with boxes and latent variables are represented by circles. Therefore the “C” represents a latent variable of class membership. In this particular example, class membership refers to the ADHD group to which each student belongs.

The model depicted in Fig. 11.1 asserts that the underlying group (i.e., class membership) is a representation of a basic truth about each person. It is a person’s membership in this group that theoretically causes their scores on the four observed items. In other words, a person who is truly hyperactive should be very likely to endorse the questions that tap hyperactive behavior. The assessed items are assumed to be related only through their common relation with the latent class variable.

The right side of Fig. 11.1 is a graphical representation of the probabilities of endorsing each of the four items. As the first two items measure inattention, responders with the inattentive subtype and the combined subtype are highly likely to endorse those items. Those with the hyperactivity subtype and those who do not have ADHD are unlikely to endorse these items. The opposite is true for the hyperactivity items. The result is four distinct profiles of response behavior, which maps to the four groups of students hypothesized to exist.

The basic structure of LCA can generalize to any behavior or characteristic including reading fluency. The four indicators from the left side of Fig. 11.1 could instead be reading fluency, reading accuracy, reading prosody, and phonological awareness, though with continuously measured variables like these, this same analysis is sometimes referred to as a latent profile analysis (LPA). If a student is truly developmentally delayed in reading, theoretically this “true risk” will cause the student to score poorly on all four reading-related tasks. Switching to the right side of Fig. 11.1, students with a global deficiency in reading would have low scores across all four reading assessments. If instead, a student has a specific deficiency in reading fluency, this specific deficit would result in a lower score for this subtest while the others remained in the relatively normal range.

## LCA Versus Factor Analysis

At this point, it is possible that LCA seems to be almost the same as a latent variable approach to factor analysis. The difference between the two types of analyses is the nature of the latent variable. In factor analysis, the latent variable is continuous, while in LCA the variable is categorical. The underlying goals of the two analyses are very similar, but while factor analysis focuses on the relations among the items, LCA focuses on the relations between the items within the person. These differences are highlighted in Table 11.1. In factor analysis, one goal is to examine the underlying dimensionality of any number of observed items to determine if they represent one factor or several factors. Similarly, LCA asks whether the observed items identify one group of responders or several groups of responders. A second goal of factor analysis is to approximate the true scores of participants on the estimated constructs, as factor analysis models the common variance across observed variables, yielding a factor score which is perfectly reliable. Instead of a true score, LCA estimates true group membership derived from probability-based classification. It is assumed that each responder's true group membership is what causes them to respond in the way they did on the observed variables. Third, factor analysis is a method of data reduction. Rather than multiple items representing a person's skills or abilities, the factor analysis provides one variable to represent the total score on several items. LCA is also a method of data reduction where the result is one variable. In the case of LCA, the one variable represents each person's assigned group membership. Rather than finding a cut score on the observed variables to place responders into groups, LCA estimates the probability that each person will belong to each group, given their responses on, and covariances among, all observed items.

In addition to the aforementioned goals, another way LCA is similar to a factor analysis is that it can be used in either a confirmatory or exploratory way. In the ADHD example, we knew we wanted to find four groups of responders; there is a clearly established history of four groups of people in respect to their ADHD diagnosis. Because of that known factor structure, we know that the LCA should identify four different groups of responders. This is not necessarily the case in many other research areas, where such criteria are unknown or unclear. When this is the

**Table 11.1** Factor analysis and LCA comparison

Shared characteristics	Factor analysis	LCA
Has a latent variable	Continuous	Categorical
Purpose 1: find the structure	Determine the underlying number of factors for all observed items	Determine the underlying number of groups in the sample of responders
Purpose 2: get the truth	Estimate the true scores of participants	Estimate the true group membership of the participants
Purpose 3: reduce the data	One variable represents scores on multiple items	One variable represents the scores and covariances of multiple items

*LCA* latent class analysis

case, LCA can be used in an exploratory way to identify whether distinct groups of people exist, how many distinct groups exist, as well as what distinguishes them. The exploratory LCA method will be demonstrated in the Application section.

## **What LCA Is Not: Cluster Analysis**

Another analytic technique that is similar to LCA is cluster analysis. Both types of analyses have the goal of identifying underlying groups of people based on their responses on the observed data and to assign those people to groups. Though LCA is similar in conceptualization to cluster analysis, it does have some key differences. First, cluster analysis estimates how a given cluster fits the observations assigned to it using the sums of squared errors approach also used in analysis of variance (ANOVA) and regression. Therefore, it falls under the same assumption rules as regression analysis in regards to the normality of the data and the importance of outliers (Cronbach & Gleser, 1953). In contrast, LCA estimates class membership based on the probability of belonging to a given group and therefore does not require the same assumptions be met. In other words, LCA is more flexible and can be used with data with non-normal distributions that demonstrate heteroskedasticity or have heterogeneity of variance. A second important difference is that LCA allows for the inclusion of all types of variables (dichotomous, categorical, count, or continuous or any combination of these) when identifying groups, while cluster analysis is limited only to continuous variables. In sum, while there are some similarities in the underlying purpose of the two techniques, LCA is a more flexible analytic tool.

## **The “So What” Section**

It is hopefully clear now that LCA can be used when the goal of the research is to identify groups. Although it is often inherently interesting to determine whether groups exist and what identifies them, this practice always begs the question: So the groups exist. So what? The next step in an LCA is to determine if the groups have some external importance or validity. Returning to the factor analysis example, this process would be akin to using the factor score of one construct to predict some other separate but related construct. Similarly, groups identified in an LCA can be examined for their relations with other constructs, either concurrently measured constructs, previously measured constructs, or as predictors of future performance.

For example, in the case of students with ADHD, the analysis identified four different groups of responders (those with an attention problem, those with a hyperactivity problem, those with a combined problem, and those with no problem). After identifying these groups, a second research question can be identified. For example, these groups of students may also have different background characteristics. Perhaps the groups of students have caregivers with different parenting styles, have different socioeconomic backgrounds, or have a family history of attention

problems. Alternatively, students in the four different groups might demonstrate significantly different academic skills, have different school attendance rates, or have different physical activity levels. In terms of predicting future performance, it may make sense to assess whether students in these groups are more or less likely to graduate from high school, whether they attend college, or even their likelihood of engaging in criminal activity. The “so what” question is reliant on the goals of the research being conducted.

## Reading and Fluency Research

Historically, LCA has been used almost exclusively in substance abuse research (e.g., Agrawal, Lynskey, Madden, Bucholz, & Heath, 2007; Muthén, 2006) and mental or psychological disorders, such as depression or attention deficit disorder (e.g., VanLang, Ferdinand, Ormel, & Verhulst, 2006). As a result, the technique is most commonly used with scores on a single test, or occasionally scores on multiple tests, with the goal of classifying individuals into known categories (e.g., the four groups of ADHD classifications or abusers vs. non-abusers). When considering this format in the context of reading research, the meaning behind the groups could take many forms depending on the observed variables included in the model. For example, current procedures for identification of a student’s risk for reading failure involve comparing student scores on a particular reading test. Rather than using a single cut point to identify those students, an LCA could be used to select students based on their scores on several different reading tests. Another way LCA could be used with reading data is to identify different profiles of strengths and weaknesses on a range of different skills. For example, if students are tested on reading decoding, reading fluency, and reading comprehension, an LCA could be used to identify whether the groups of students existed with relative strengths or weaknesses in each of these areas, and could then be targeted for intervention in those areas of need or enhanced instruction. Though thousands of articles exist that examine children’s reading development using reading fluency and curriculum-based measurement, few have done so using LCA.

Because so few examples of research questions regarding reading fluency have been addressed using LCA, the remainder of this section includes some examples from the broader base of education research that make use of this technique, each of which uses reading fluency in some way. We focus on ideas of how educational research questions can be framed and answered using LCA. Three studies will be examined in depth; one examining adult dyslexic readers, one examining school-level risk, and a third examining adolescent reading skills. These three disparate examples are given to highlight the flexibility of LCA and the value of the person centered approach. This is done in the hope that you will be more readily able to map on your specific research question to the LCA technique.

### ***Example 1: Adult Dyslexic Readers***

Though LCA examples in reading fluency are scarce, some work in this area has been done by Leinonen and colleagues (Leinonen, Müller, Leppänen, Aro, Ahonen, & Lyytinen, 2001). In this article, the researchers used cluster analysis to identify whether groups of adult dyslexic readers could be identified based on their reading accuracy and reading speed. The results of the cluster analysis identified four profiles of dyslexic readers. The first identified group was the “hasty dyslexic readers” group, who read quickly but made a lot of errors. The second group, hesitant dyslexic readers, made a few errors but read slowly. Third, a group of mildly dyslexic readers was identified, who were relatively fast and made relatively few errors compared to the rest of the dyslexic sample. Finally, the fourth group demonstrated extremely high error rates and very slow reading speed, and was named the severe dyslexic group. The researchers next examined whether the four groups identified in the cluster analysis would be significantly different from a normative sample on measures of phonological processes. Though the Leinonen et al. (2001) paper used cluster analysis, the same questions could be addressed with LCA: identifying groups of dyslexic readers based on their text accuracy and text reading speed.

### ***Example 2: School Risk***

It is important to note that the analytic technique does not have to be used exclusively to identify groups of *people*. Logan and Petscher (2010) used LCA to identify groups of schools based on the percentages of at-risk students the school served. Historically, schools with higher percentages of students considered to be at risk—based on their language, minority status, and poverty status—are found to also perform more poorly on state standardized tests. Previous research in the area of school risk used an arbitrary cut point on only one of the aforementioned risk factors, such that schools above the cut point were deemed at-risk, while those below the cut point were considered not-at-risk.

Logan and Petscher assessed the percentage of English language learners, minority students, and students eligible for free or reduced-priced lunch in a large sample of elementary schools ( $n=569$ ) to determine whether distinct groups of schools existed based on these factors (2010). They found four groups of schools. Similar to the previous example, two classifications of schools followed a typical “high vs. low risk” pattern: one group of schools had very low percentage of English language learners, minority students, and students living in poverty, and one had very high percentages of those same students. The remaining two groups had relatively high percentage of students living in poverty and of minority status, but one had high percentage of English language learners, while one did not.

Next, for the “so what” portion of the analysis, the authors wanted to determine if the students had different reading skills depending on their school’s assigned group membership from the LCA. In this study, the researchers had measured stu-

dents four times during the year, which allowed for growth models to be fit to the data. The school clusters extracted from the LCA were added to the growth model as predictors. This was done in such a way that separate estimates were obtained for each identified cluster. Because four clusters of schools were identified by the analysis, four separate sets of mean growth rates and beginning-of-year intercepts, were obtained—one for each group. The authors found significant differences between the four groups in the intercepts (where students began the year) and slopes (how quickly they grew during the year). This validated the theory that multiple indicators of school risk should be considered.

### ***Example 3: Adolescent Reading Ability***

Although much focus has been given to students' reading skills early in their schooling, recently there has been a shift of this focus to the issues facing older students who are struggling to read, particularly those who struggle with reading comprehension. In a 2011 paper, researchers sought to identify the groups of adolescent students (ninth grade) based on their reading comprehension skills (Brasseur-Hock, Hock, Kieffer, Biancarosa, & Deshler, 2011). Brasseur-Hock and colleagues used three different assessments of reading comprehension: the Kansas state-administered Reading Assessment, the WLPB-R Passage Comprehension subtest (Woodcock, 1991), and the GORT-4 Comprehension subtest (Wiederholt & Bryant, 2001); and identified four different profiles of students: struggling comprehenders, low-average comprehenders, average comprehenders, and advanced comprehenders.

The identified reading comprehension groups have some inherent importance, but Brasseur-Hock and colleagues took the analysis a step further. They examined reading fluency as a part of the “so what” portion of this study. The researchers conducted a second LCA based on the component skills of only those students in the lowest two groups of reading comprehension (as identified in the first LCA). The goal of this work was to determine if different groups of poor comprehenders could be identified based on their reading fluency, reading accuracy, and language comprehension. The results of this LCA identified five groups of poor comprehenders: global weakness, moderate global weakness, weak language comprehension, weak reading comprehension, and germane to the present focus, they also identified a group of dysfluent readers (who read more slowly than their other poor-comprehending peers). These different profiles suggest that reading comprehension is complex, with many underlying causes and therefore many different approaches to potential treatments.

These three examples demonstrate one other unique aspect of LCA. In each example, the groups identified in an LCA provide more information than you could get using all of the observed variables in a regression-based statistical test (like *t*-test, ANOVA, or multiple regression). Sometimes called variable-centered approaches, a regression-based test would identify how much of the variance in the outcome can be uniquely explained by each of the observed variables. In contrast, an LCA independently considers the relative position of each person on a whole set of predictors. The resulting latent class variable encompasses how each person



scores on several different items, as well as how their scores on those items *covary*. In other words, LCA provides a way of capturing the variability within a person's scores, and is for that reason sometimes referred to as a person-centered approach.

## When Do I Use It?

When a researcher begins to conceptualize a study that will use LCA, there are several important considerations. A summary of these has been included in the LCA guidelines callout box below. Generally, LCA should be used when you have research questions about identifying groups of responders (e.g., people, classrooms, schools, parents), whether the observed variables are categorical or continuous. The term LPA can be used when the observed variables are continuous (e.g., Muthén & Muthén, 2000). But as the conceptualization is identical, LCA is an overarching term that can be used to refer to both models. It is important when considering this analysis to conceptualize how you intend to validate the results. Research involving an LCA should have a two-part question. What follows is an in-depth description of the research question that will be used to guide the analyses conducted during the remainder of the chapter. Notice that the relations among the constructs have been written in the language of both a variable-centered and a person-centered approach.

### Some LCA Guidelines

As you are planning to use this analysis to answer your own research questions, you should take the following into consideration:

1. **Use two-part questions:** Identifying groups in your sample is only the first step. Think through what you plan to do with the groups. In other words, decide on your “so what” question.
2. **More is better:** The more observed variables you have, the more stable the model is, and the more likely you are to reach convergence. We recommend three as the minimum number of observed variables (Goodman, 1974).
3. **But not too many more:** More observed variables require more observations (a larger sample size) in order to find a solution. Be cautious of your sample size when deciding how many observed variables to use in the LCA. Simulation studies are useful to determine the sample size you need to measure the effect you are interested in.
4. **Constructs matter:** The types of groups LCA identifies depend on the observed variables that go into it. If the observed variables measure all the same theoretical construct, the LCA can be considered an alternative to a quantitative difference, or a cut point. If the observed variables measure different constructs, then the LCA will identify *qualitative* differences between groups on the constructs of interest.

5. **Use similar scales:** If the observed variables are on different scales (e.g., if one has a mean of 100 and standard deviation of 15, and another has a mean of 10 and standard deviation of 0.40), the model may not fit easily. When this is the case, use a standardization procedure to convert the observed variables to  $z$ -scores. Like cluster analysis, standardizing data for LCA can help with interpretation, but unlike cluster analysis, LCA clustering is invariant to linear transformations of data.

## Application to Reading Fluency

### *Research Question*

The focus of the remainder of the chapter is students' early reading skills. Reading comprehension is a critical skill in early elementary, as reading is required to learn about all other school topics and is especially critical as students move from "learning to read" to "reading to learn" (Adams, 1990). Curriculum-based measurement of reading fluency has often been demonstrated to be a good predictor of later reading comprehension in variable-centered approaches (e.g., Fuchs, Fuchs, & Maxwell, 1998; Gough, Hoover, & Peterson, 1996). Variable-centered approaches to this problem typically ask a research question such as, "Can we predict reading comprehension at the end of Grade 2 from student's beginning of school-year scores on several reading fluency assessments?" However, we argue that it is important to understand if there are reliable profiles of students on reading fluency, because these can provide a data-based guideline for identifying which students are at risk for later reading comprehension difficulties, necessitating a person-centered approach. The person-centered approach to this research question will be examined for the remainder of this chapter, and has two parts. First, we will determine if there are reliable profiles of students based on their reading fluency skills at the beginning of Grade 1, and second, we will determine whether students in these profiles demonstrate significantly different scores on reading comprehension at the end of Grade 2.

### *Sample*

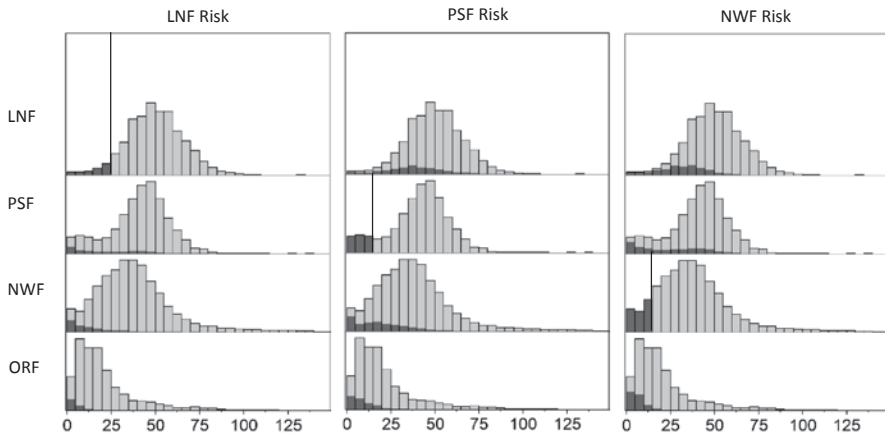
The sample was drawn from the Progress Monitoring and Reporting Network (PMRN), an archival data set containing data on students in every Reading First school in the state of Florida. The present sample included 17,830 students who were followed longitudinally at six assessment points through first and second grade in the 2004–2005 and 2005–2006 school years. These students were enrolled

in 342 different schools across 37 districts throughout the state of Florida. The sample of children was 1.4% Asian, 31.4% Black, 22.5% Hispanic, 4.4% Multiracial, and 39% White. Regarding their language proficiency, 82.9% were native English speakers (5.7% graduated, 10.7% still enrolled in an ESOL program). The sample had relatively high poverty, with 72% of students eligible for free or reduced-priced lunch. Finally, a small percentage of students were identified with a primary exceptionality: 6.8% had speech impairment, 2.4% had language impairment, 2.2% had a specific learning disability, and 1% of the sample was considered gifted.

## Measures

In the present application, reading comprehension skills were assessed at the end of second grade with the Stanford Achievement Test, 10th edition (SAT 10; Harcourt Brace, 2003). The SAT 10 is a widely used standardized measure administered to students in a group setting by the classroom teacher. In the test, students read passages and answer multiple-choice questions about the content. The SAT 10 measure will be used in the “so what” portion of the analysis. The assessments included in the LCA in the present study are four subtests of the *Dynamic Indicators of Basic Early Literacy (DIBELS)*; 5th edition; Good et al., 2001): phonemic segmentation fluency (PSF), letter naming fluency (LNF), nonsense word fluency (NWF), and ORF. For each subtest, the administrator noted the number of errors and reported the number of stimuli read correctly per minute (higher is better). In LNF, students were asked to read presented letters as quickly as possible. In the PSF subtest, students were required to provide the individual phonemes from a given word (e.g., say all the sounds in “ ”). NWF is a nonword reading task (e.g., say all the sounds in “mip”). Finally, ORF performance was measured by having students read three separate passages aloud, with final scores representing the median fluency (Good et al., 2001, p. 30). The included data set contains the four DIBELS subtests, the SAT 10, and a variable representing the child’s ID.

Each of the four DIBELS subtests represents one of four different skills that underlie reading ability. The four skills are specific but similar; they are related but not perfectly correlated. To illustrate, the distributions of each skill have been graphed in Fig. 11.2. There are three columns (panels); each with four rows. Each row represents a different reading fluency skill. The first column (or panel) highlights those students who are considered at risk in LNF based on the DIBELS benchmark. In that column, the same students’ scores on the three other administered subtests have also been highlighted. Note that the students deemed to be “at risk” in LNF have scores across the distribution of PSF and NWF. The same is true for cutoffs on the three other subtests. Thus, it is possible that a student who has difficulty in one skill may be proficient in another. The heterogeneity and covariance among the four indicators would be ignored in a variable-centered approach like ANOVA, regression, or factor analysis, but can be captured in an LCA.



**Fig. 11.2** Histogram of the distributions of the four examined reading fluency subtests

### *Steps in LCA*

An LCA requires that the researcher provides an estimated number of classes to extract. When the true number of underlying groups is not predetermined, an exploratory analysis can be conducted to determine the optimal number of groups for the data (see Muthén, 2008). The exploratory approach requires that several models be fit to the data, each with an increasing number of groups. Model comparisons (discussed in subsequent sections) are then used to determine which of the models, and thus which number of classes, is the best fit to the data. Based on the specified number of groups, the LCA will provide a suggested group membership for each person. For example, if LCA was used to cluster the students into two groups, it would likely find one low group (students with poor fluency) and one high group (students with good fluency skills). Usually, in an exploratory analysis we will fit six models to the data, including between two and seven groups, though more groups can be fit if the data or hypotheses call for it.

To place students in groups, LCA calculates the probabilities that each child belongs to each of the extracted groups, based on the child's responses on the observed items. The probabilities take into account both the value (high or low) of a child's response on each item, as well as how similar or different the responses are to each other (the covariance among the different responses within person). The analyses examine each student's probabilities of belonging to each group, and assign him or her to the group with the highest probability. If the LCA is asked to extract three groups, students could theoretically have a 33% chance of belonging to the first two groups, and a 34% chance of belonging to a third group. In this case, the student would be assigned to the third group even though the probability of his belonging there is not very high. In this way, LCA is fundamentally different from the cut score or threshold approach of making groups. LCA does not attempt to identify a

cut score or threshold to the observed response or pattern of responses. In fact, it is possible that two people who have exactly the same score on one response variable can be placed in different groups, if their scores on the other items are different.

### Getting Started with Mplus

Though a complete description of Mplus is beyond the scope of this particular chapter, we will provide with some basic information to help you get started. First, the data file should be saved as a text file. Two data files are provided with this chapter; one in SPSS and the same in a tab-delimited text file. It is the tab-delimited file that Mplus uses to run analyses. For an introduction to working with Mplus, the program, see the introductory text *Data Analysis with Mplus* by Christian Geiser (2012). Figure 11.3 contains the input text to run a two-group LCA. There are several pieces of code that are important to working with Mplus, and all are described thoroughly in the Mplus user guide.

Briefly, the DATA command lists the name of the data file and describes the format. The VARIABLE command lists the names of all variables in the data file, and selects which ones will be used for the particular analysis. Because the goal of this particular analysis is to determine whether there are identifiable groups based on the students' scores over the year, the USEVARIABLES command contains the variable names of the four assessments of ORF in Grade 1. The MISSING command lets you indicate how missing data are coded. In our case missing data is blank. Finally, the piece of code that tells Mplus that you want to run an LCA is the "CLASSES"

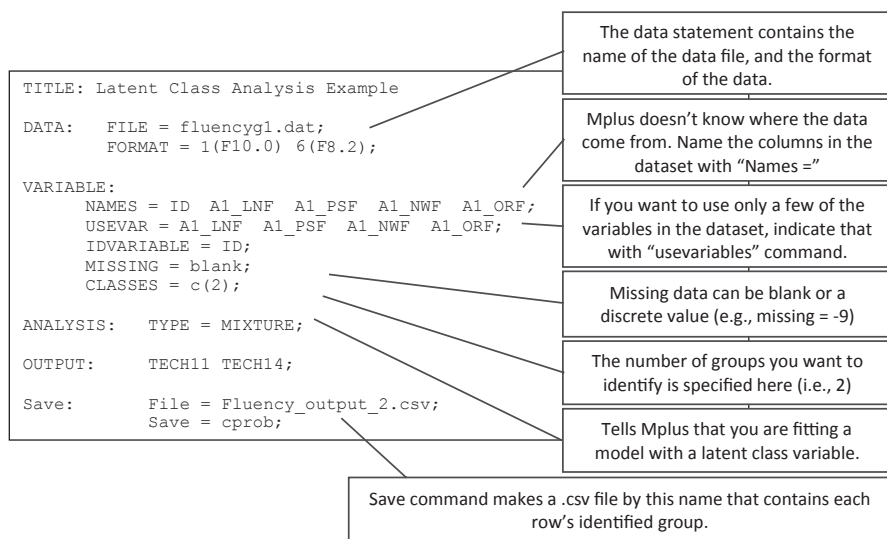


Fig. 11.3 Annotated Mplus input for two-group model

line. The letter “c” here represents the name of the latent class variable. The number in parentheses is the number of groups you want the analysis to find. Here, we are asking for two. When you increase the number of groups, you can save the input file with a new name, change the number of classes as indicated on this line, and rerun the analysis. When this program is run, Mplus automatically opens the almost identical looking output file. There are a few pieces of output you will examine closely for indication of model fit, the locations of which are explained here.

## *Selecting the Final Model*

As previously indicated, when conducting an LCA, several models will be run in the course of analysis. The first model that the researcher actively fits is the two-group model, after which several additional models are run, each increasing the number of groups by one. Deciding on the final model to report is a balancing act between several model fit indices and the interpretability of the results. When considering that decision, we discuss model fit first, followed by interpretability and decision-making.

### **Model Fit**

When evaluating model fit, it is necessary to examine and balance several different criteria. A few of the statistical tests are automatically calculated in Mplus when conducting an LCA. These statistics can be found in the output. Directly after the information about the run of the model there is a section of the output called “Model Fit Information.” The number of free parameters is the first piece of information in this section. The  $-2LL$  information is contained here (labeled Loglikelihood, H0 value). Both of these can be harvested for a chi-squared comparison test between models. Other statistics reported in this section include the Bayesian information criterion (BIC; Schwarz, 1978) and the Akaike information criterion (AIC; Akaike, 1974), which are used to compare relative model fit. The AIC and BIC are popular measures from the structural equation modeling literature. They combine fit and complexity to compare model parsimony, with lower values indicating better model fit. Model fit for BIC and AIC are only useful in relative terms. If the ratio of the BIC for model A (model with fewer groups) to the BIC for model B (model with more groups) will typically be  $>1$ , suggesting that the model with more groups is a better fit. One rule of thumb suggests that if the value of the ratio is between 1 and 3, it indicates minimal evidence, between 3 and 10 indicates some evidence, between 10 and 100 is strong evidence, and greater than 100 indicates decisive evidence of an increase in model fit (Raftery, 1995). These values are extracted from each of the models run, and are typically plotted to examine their incremental change with each additional group added to the model. When the slope of the plotted AIC or BIC curve begins to flatten, it is an indication that there is very little information gained relative to the number of degrees of freedom sacrificed for the model to identify additional groups.

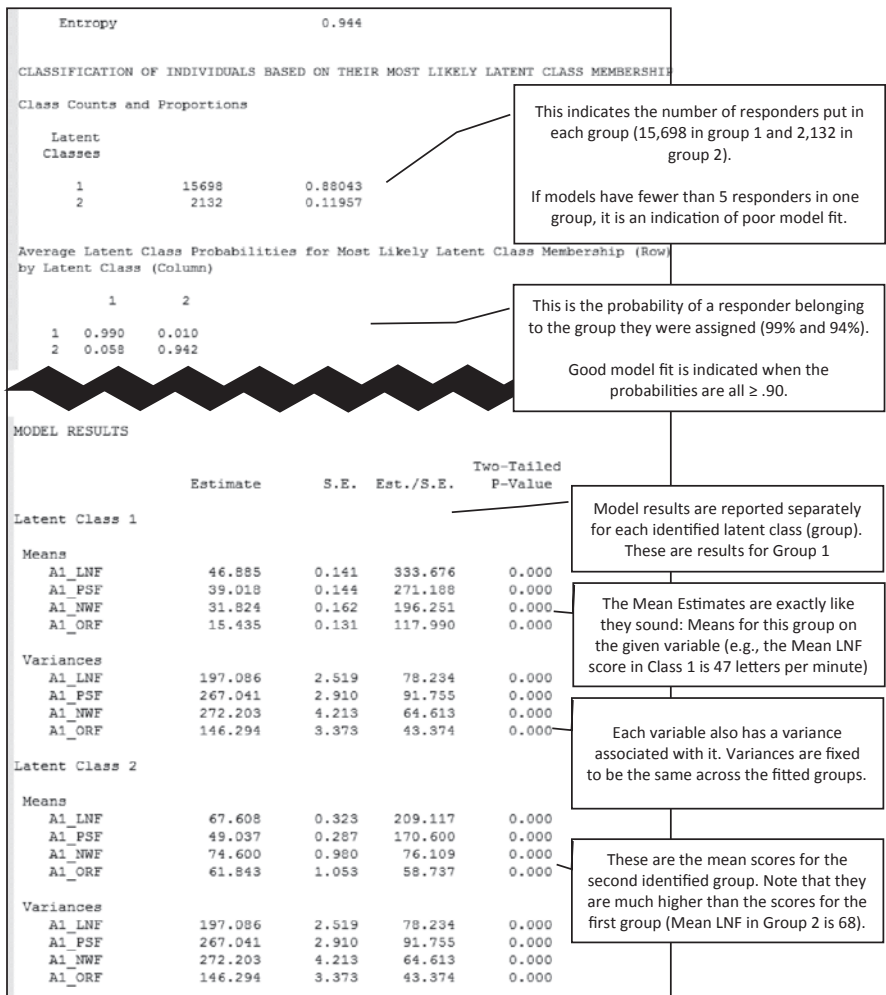


Fig. 11.4 Annotated Mplus output for two-group model

A second statistical test automatically calculated in an LCA conducted in Mplus is entropy (Ramaswamy, DeSarbo, Reibstein, & Robinson, 1993). The entropy statistic can be used to determine how separated the identified groups are from one another. In other words, this statistic can measure how much differentiation there is between the different groups. Entropy values greater than .80 indicate a good separation of the identified groups (Ramaswamy et al., 1993). Entropy is also automatically reported in Mplus, is clearly labeled, and is displayed in the output presented in Fig. 11.4. There are also two tests that can be estimated in Mplus if they are specifically asked for in the input file (See Fig. 11.3, the line including Tech 11 and Tech 14). These are the Lo–Mendell–Rubin likelihood ratio test (specified as TECH 11; Lo, Mendell, & Rubin, 2001) and a parametric bootstrapped likelihood



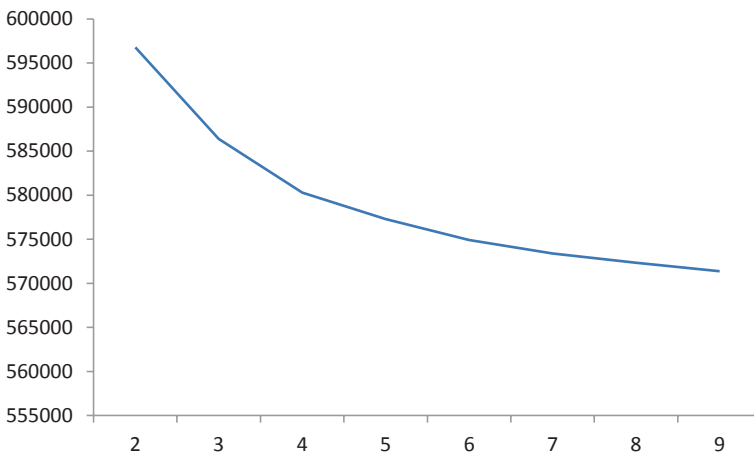
**Table 11.2** Model fit indices for each model fit to the fluency data

Number of groups	-2LL	Free parameters	AIC	BIC	Entropy	Tech 11	Tech 14
2	-298375	13	596775.3	596876.5	0.944	0	0
3	-293178	18	586392.7	586532.9	0.793	0	0
4	-290117	23	580279.4	580458.5	0.834	0	0
5	-288609	28	577273.9	577492	0.805	0.014	0
6	-287420	33	574905	575162.1	0.819	0.0016	0
7	-286660	38	573395.8	573691.7	0.834	0.0361	0
8	-286121	43	572328.9	572663.8	0.844	0.025	0
9	-285635	48	571366.7	571740.5	0.846	0.6802	0

*AIC* Akaike information criterion, *BIC* Bayesian information criterion

ratio test (specified as TECH 14; McLachlan & Peel, 2000). The null hypothesis of these statistical tests is that the model being tested has identical model fit to a model with one less group (e.g., a model with four groups fits the same as a model with three groups). When the *p*-value is significant, it indicates that the null hypothesis is rejected, and the current model fits better than the previous one.

To conduct these analyses, it is necessary to extract this information from the output of each model that is run. In the current application eight models were run, each testing for a different number of groups. We have extracted the model fit indices from each one of the outputs and placed them in Table 11.2. The entropy statistic looks good for all fitted models (values  $\geq .80$ ). The Tech 14 did not show any change, but the nonsignificant *p*-value for Tech 11 in the nine-class model indicates that the nine-class model does not fit better than the eight-class model. Both the AIC and BIC are evaluated in relative terms, and both decrease as more groups are added to the model. To assess relative model fit, we plotted the values of the AIC, which are presented in Fig. 11.5. In Fig. 11.5, there is a sharp decline from the two-group model to the four-group model, a slightly less steep decline from the four-group to the five-group model, and an even weaker decline after that point. This is where the



**Fig. 11.5** AIC values from models with 2–9 groups

balancing act begins. Each of these indices suggests a different number of groups should be retained. We tend to lean toward parsimony, thus selecting either the four-group or five-group model to be retained.

### Model Interpretability

As indicated in the previous section, the decision of which model to retain also involves how interpretable the identified models are. This step involves two main indexes. The first is the number of responders placed in each group (See Fig. 11.4). If provided the opportunity, an LCA could potentially place individual responders in their own group. But this does not make conceptual sense because one person does not make a group. A good conceptual rule of thumb is that each identified class (group) should contain at least 5% of responders, or at least five responders in our case (based on simulations presented by Nylund, Asparouhov, & Muthén, 2007). If a group has fewer than five people, it is more difficult to ascertain whether the participants represent a true group of responders rather than a chance group of outlying responders. If the model fit indices compared in the first step suggest a model be retained even when one group has fewer than five responders, you should instead select a model with fewer identified groups.

The next, and arguably most important, indicator of model fit is whether the identified groups make theoretical sense. As indicated in Fig. 11.4, LCA provides mean scores for each identified group. By examining the mean scores on each group, we can identify what the groups conceptually represent. For example, in the two-group model presented in Fig. 11.4, there is one group with relatively low scores and one group with relatively high scores. All identified means are within the range of observed data. In addition, the groups make theoretical sense; if one group was identified containing hundreds of students with very poor LNF skills (the most basic subtest) but very high ORF skills (the most advanced subtest), this would not make sense given the visual representation of the data in Fig. 11.2.

### Decision-Making

The examination of the aforementioned fit indices culminates in a decision of which model to retain. At this point in the analysis, the decision becomes a balancing act. Several fit indices have been discussed, and these may provide contradicting information to one another. Each of the factors should be considered when making the final decision.

First, the AIC/BIC graph should be examined to give a general idea of a more narrow range of potential final solutions. Based on the previously described analysis of the graph (Fig. 11.5), the model fit indices extracted from the results from our series of analyses suggest that we could select either the four-group model or the five-group model. Next, return to Table 11.2 and examine whether entropy is sufficiently large to indicate good fit. Entropy is  $>.80$  for both the four-class and five-class models, indicating good model fit. Next, examine the Tech 11 or Tech 14 outputs to determine if they also agree with the conclusion that either the four-class or five-class model is most appropriate (Table 11.2). In this case, they do not. The Tech 14

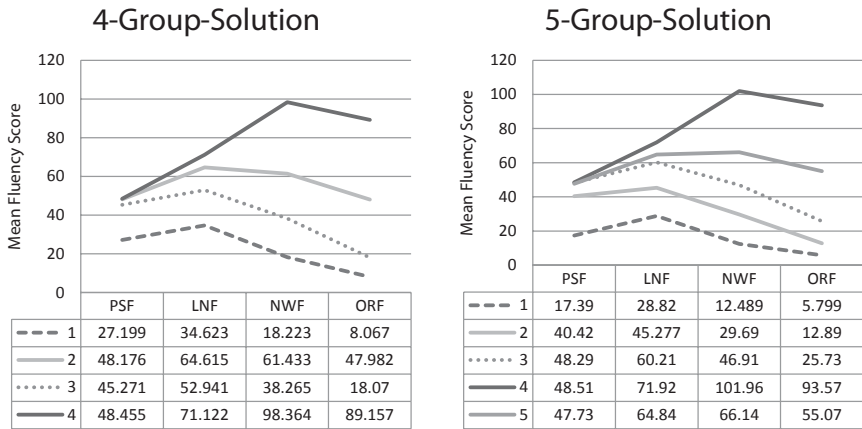


Fig. 11.6 Four- and five-group solutions

option is never larger than zero, while Tech 11 suggests that model fit improves with each increasing number of classes, except the model with nine groups does not fit better than eight. Based on the BIC graph, we know that eight is not a reasonable number of groups. This leaves us still considering the differences between the four- and five-group models. Next, we can move to determining whether the models make theoretical sense. To aid with the decision of which model to retain, we have plotted the estimated mean values for each of the extracted groups in the two front-running models in separate graphs (Fig. 11.6).

At this point, it is a good idea to give the groups names to aid in the decision about which to keep. In the case of the four DIBELS subtests, these have been normed such that we have suggested benchmarks indicating on track student progress at the beginning of Grade 1; these are a score of 35 for PSF, 37 for LNF, 24 for NWF, and 13 for ORF. Given these statistics, the four groups estimated in the four-class solution could be named as follows: (1) low performers, (2) average performers, (3) above average performers, and (4) advanced performers. The five-class solution identifies these same groups, but also identifies a group best conceptualized as “slightly above average performers” (group 3 in the second panel of Fig. 11.6). This group is slightly above the “average performers” group, and below the “above average performers” group. We find the “slightly above average performers” group to be lacking specificity; there is nothing about the “slightly above average” group that makes it unique from either of the surrounding groups. This means we have finally arrived at the decision of the number of groups represented by the data, and will retain the four-group model.

### Reviewing the Model Fit Indices

When conducting an LCA, there are five indicators of model fit that should be balanced to determine the final model: (1) the AIC/BIC, which in this case indicated either the four- or five-class model be retained, (2) the  $-2LL$  and number of free

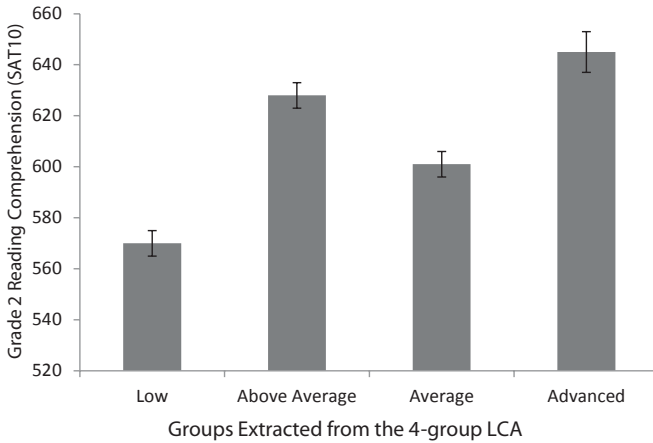
parameters, (3) the Tech 11 and Tech 14 outputs, both of which suggested each model fit is better than the previous one, (4) model entropy and (5) the number of people placed in each group, both of which were sufficient for all models fitted, and finally (6) that the groups identified by the model are theoretically sound. Balancing all five fit indicators is where LCA becomes more like an art than a science. Based on balancing all of these fit indices, we identified the four-group model as the best fit to the data.

The first research question has now been answered: Yes, reliable groups of students were identified based on their Grade 1 fluency scores. The students fall into the groups we have termed as low performers, average performers, above average performers, and advanced performers. Notably, no groups emerged suggesting relative weaknesses in one area of fluency-based reading skill, rather the differences were more quantitative in nature. It is important to note that the PSF subtest seems to hit a ceiling; all three of the higher groups have approximately the same score on this subtest. The greatest differentiation between groups is observed in the ORF subtest.

## So What?

The results of the first research question suggest that the best fit to the data is a model with four groups. Now we can move to the second research question: determining the extent to which membership in one of these four groups is related to later reading comprehension skills. To do so, we need to know to which group each child belongs. This step is accomplished by saving a data file that includes this information along with each child's ID. This can be requested in the Mplus input program with the "save" command line (See Fig. 11.3). The first line of this command tells Mplus what to name the file. When naming the file, we will typically include the number of groups fitted in that particular model (note the name Fluency\_output\_2). The second line (save = cprob) tells Mplus to write the probability of each student belonging to a specific group, along with their assigned group to the data file. Next, the data file that contained each student's assigned group membership from the four-group solution was merged into the original SPSS database. Then the variable of class membership is able to be used as a predictor or an outcome in any other analyses.

For our purposes, the "so what" question asked whether student reading comprehension scores varied as a function of the groups we identified in the LCA. This could be done in a number of ways, but one of the simplest is to predict later reading comprehension scores from the group membership variable extracted from Mplus. This analysis was conducted in SPSS using the General Linear Model procedure. We requested that the program provide Bonferroni-corrected post hoc contrasts to test all pair-wise comparisons between the four groups. The post hoc contrasts test whether the mean reading comprehension score for each group identified by the LCA was significantly different from each other group. Figure 11.7 contains the results of the generalized linear model (GLM) with post hoc contrasts, the four groups are listed in the order of their extraction, but with the names assigned to them during the final LCA decision-making process. All contrasts were significant ( $p < .001$ ), suggesting that each group identified in the LCA had a significantly different score



**Fig. 11.7** Reading comprehension scores for the four groups extracted by the LCA

on reading comprehension. Cohen’s *d* effect sizes were calculated for the difference between each pair of groups, and those results are reported in the above diagonal portion of Table 11.3. The smallest effect size was .55, which is considered a medium effect and corresponds to the one half standard deviation difference between the above average performers and advanced performers groups (Cohen, 1988). All other effect sizes were above the suggested benchmark for a large effect (Table 11.3).

The results of the GLM suggest that different profiles of fluency abilities at the beginning of Grade 1 predict later reading comprehension. Our findings converge with theoretical work hypothesizing that ORF may serve as an indicator of overall reading competence (e.g., Le Berge & Samuels, 1974), as well as the body of empirical work demonstrating the predictive relation between reading fluency and comprehension (e.g., Gough, Hoover and Peterson, 1996). Given that reading comprehension is a skill relied upon for all other academic skills in elementary school and beyond, understanding indicators of this skill (such as reading fluency) are of great impor-

**Table 11.3** Differences in reading comprehension scores between the four groups extracted from the LCA

	Low	Above Average	Average	Advanced
Low	572.34 (35.23)	-1.79	-0.91	-2.3
Above Average	-58.97	631.31 (30.63)	0.87	-0.55
Average	-31.21	27.76	603.55 (33.35)	-1.39
Advanced	-75.74	-16.77	-44.53	648.08 (30.63)

*Note:* Means (standard deviations) for each group on the diagonal. Below diagonal are calculated mean differences, and above diagonal are Cohen’s *d* effect sizes (in each case, column is subtracted from row)

tance. Utilizing LCA in such endeavors, as we have done in this chapter, allows us a more nuanced understanding of the relation between different indicators of fluency and reading comprehension than variable-based methods—this is especially important given propositions that reading fluency represents a complicated, multifaceted construct (Fuchs, Fuchs, Hosp, & Jenkins, 2001). In the current study, we used four different measures of fluency, classifying students based on how they covaried on all four skills simultaneously to get an estimated group membership. Had we only used one of our four indicators of reading fluency to investigate the relation between fluency and reading comprehension, we would have been provided with information on individual risk or performance relevant only to one aspect of reading fluency. Based on Fig. 11.3, it is clear that if any one indicator had been selected it would likely exclude students who struggle in a different area. Additionally, LCA provides benefits beyond the use of an overall fluency score, such that it allows us to also look at the covariance among the four indicators, without using a cut score.

## **LCA Considerations and Extensions**

### ***Sample Size Recommendations***

A good rule of thumb for sample size in LCA is a sample size of at least 250 unique people, places, or things should be observed to begin to approach reliability. Still, some recommendations are as high as 500 unique observations to constitute best practices (Nylund et al., 2007). It is important to consider that LCA is essentially a factor analysis, and thus is subject to potential variation in the reliability of the estimate depending on the reliability of the constructs being measured, the means of the observed constructs, and the covariance among the observed constructs.

The necessary sample size will also depend on the number of observed variables included in the LCA. More variables mean more potential solutions will be fit to the data, which can increase the needed sample size to achieve good model performance. This problem is compounded when the observed variables are continuous; continuous variables have more potential solutions than do categorical ones. Use caution when deciding to increase the number of observed variables in an LCA. Conducting data simulations is the most cautious and most effective way to estimate the number of responders needed to detect a specific effect.

### ***Potential Problems***

As hinted at in the previous paragraph, LCA can be a fickle analysis to fit. Errors in Mplus are indicated in the output, immediately following the iteration history. It is important to check for errors as they may indicate that the model has not achieved a viable solution. One potential error you may encounter is the idea of a local maximum. You can think of local maximum as a fake solution.

Because LCA is a factor analysis, it is subject to the same problems that factor analysis is, including method variance, rater bias, and historical effects. For example, imagine an LCA containing five observed variables: two observed variables that share rater bias (e.g., two rating scales filled out by the teacher) and three other variables that do not (e.g., measured by different trained graduate students). In this situation, the LCA may be likely to form clusters that always covary on the two observed variables that share the rater bias, merely because of the rater bias rather than the actual construct being assessed. Consideration of these biases should play a role in the decision of which variables to include as well as the interpretation of the groups identified in the LCA. As a result, it is a good practice to eliminate variables that contain these biases from inclusion in the model.

### *Extensions*

In addition to the ways it has been used here, LCA can also be extended when combined with other techniques. First, LCA has an underlying assumption that all observed variables are uncorrelated within a class (i.e., the correlations are set to zero). The factor mixture model (e.g., Lubke & Muthén, 2005) relaxes that assumption. By simultaneously fitting an LCA and a factor analysis, the observed variables are allowed to be related through a latent factor in addition to the latent class variable. Factor mixture analysis can also be extended to a multilevel model (e.g., Torppa et al., 2007). Another extension of the basic LCA model is the growth mixture model (e.g., Kreisman, 2003), which combines LCA with a growth model. In these models, latent classes are estimated based on the responders' rank order at the intercept, as well as how quickly they change during the measured time periods. LCA can also be used over time to determine whether responders stay in the same groups over repeated measurements. This extension is called the latent transition analysis (e.g., Compton, Fuchs, Fuchs, Elleman, & Gilbert, 2008; Ding, Richardson, & Schnell, 2013).

### *Conclusion*

This chapter introduced the concept of LCA and provided a demonstration of it in the context of reading fluency. The LCA technique has been slow to be integrated into education research, despite the match between the needs of the education research literature and the solution that LCA provides. This may be because most research using the technique has been conducted in other fields, perhaps masking the potential benefits of the technique to those in education and reading fluency research. By providing examples of the many uses that LCA can have in education and specifically fluency research, examples of reframing variable-centered questions to person-centered questions, and the walk through of the actual process of fitting a LCA model, we hope that researchers in the area of reading fluency research can begin to incorporate the LCA into their research.



## References

- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: The MIT Press.
- Agrawal, A., Lynskey, M. T., Madden, P. A., Bucholz, K. K., & Heath, A. C. (2007). A latent class analysis of illicit drug abuse/dependence: results from the National Epidemiological Survey on Alcohol and Related Conditions. *Addiction*, *102*(1), 94–104.
- Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.
- Brasseur-Hock, I. F., Hock, M. F., Kieffer, M. J., Biancarosa, G., & Deshler, D. D. (2011). Adolescent struggling readers in urban schools: results of a latent class analysis. *Learning and Individual Differences*, *21*(4), 438–452.
- Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin*, *50*(6), 456.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates.
- Compton, D. L., Fuchs, D., Fuchs, L. S., Elleman, A. M., & Gilbert, J. K. (2008). Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences*, *18*(3), 329–337.
- Ding, C., Richardson, L., & Schnell, T. (2013). A developmental perspective on word literacy from kindergarten through the second grade. *The Journal of Educational Research*, *106*(2), 132–145.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific studies of reading*, *5*(3), 239–256.
- Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal measures of reading comprehension. *Remedial and Special Education*, *9*(2), 20–28.
- Geiser, C. (2012). *Data analysis with Mplus*. Guilford Press.
- Good, R. H., Kaminski, R. A., Smith, S., Laimon, D., & Dill, S. (2003). *Dynamic indicators of basic early literacy skills*. Longmont, Colorado: Sopris West Educational Services.
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.
- Gough, P. B., Hoover, W., & Peterson, C. L. (1996). Some observations on the simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties* (pp. 1–13). Mahwah: Lawrence Erlbaum Associates, Inc.
- Harcourt Brace (2003). *Stanford Achievement Test* (10th ed.). San Antonio: Author.
- Hudziak, J.J., Heath, A.C., Madden, P.F., Reich, W., Bucholz, K.K., Slutske, W., Bierut, L.J., Neuman, R.J., & Todd, R.D. (1998). Latent class and factor analysis of DSM-IV ADHD: A twin study of female adolescents. *Journal of American Academy of Child and Adolescent Psychiatry*, *37*, 848–857.
- Kreisman, M. B. (2003). Evaluating academic outcomes of Head Start: An application of general growth mixture modeling. *Early Childhood Research Quarterly*, *18*(2), 238–254.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, *6*(2), 293–323.
- van Lang, N. D., Ferdinand, R. F., Ormel, J., & Verhulst, F. C. (2006). Latent class analysis of anxiety and depressive symptoms of the Youth Self-Report in a general population sample of young adolescents. *Behaviour research and therapy*, *44*(6), 849–860.
- Leinonen, S., Müller, K., Leppänen, P. H., Aro, M., Ahonen, T., & Lyytinen, H. (2001). Heterogeneity in adult dyslexic readers: Relating processing skills to the speed and accuracy of oral text reading. *Reading and Writing*, *14*, 265–296.
- Lo, Y. Mendell, N. R., & Rubin, D.B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778. doi:10.1093/biomet/88.3.767.
- Logan, J.A.R., & Petscher, Y. (2010). School profiles of at-risk student concentration: Differential growth in oral reading fluency. *Journal of school psychology*, *48*(2), 163–186.

- Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods, 10*(1), 21.
- McLachlan, G., & Peel, D. (2000). Mixtures of factor analyzers. *Finite Mixture Models, 238–256*.
- Muthén, B. (2006). Should substance use disorders be considered as categorical or dimensional? *Addiction, 101*(Suppl. 1), 6–16.
- Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models*. Charlotte: Information Age Publishing.
- Muthén, L. K., & Muthén, B. O. (1998–2006). *Mplus user's guide*. Los Angeles: Authors.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling, 14*(4), 535–569.
- Rasmussen, E. R., Neuman, R. J., Heath, A. C., Levy, F., Hay, D. A., & Todd, R. D. (2002). Replication of the latent class structure of Attention—Deficit/Hyperactivity Disorder (ADHD) subtypes in a sample of Australian twins. *Journal of Child Psychology and Psychiatry, 43*(8), 1018–1028.
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology, 25*, 111–164.
- Ramaswamy, V., DeSarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science, 12*(1), 103–124. doi:10.1287/mksc.12.1.103.
- Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6*(2), 461–464.
- Torppa, M., Tolvanen, A., Poikkeus, A. M., Eklund, K., Lerkkanen, M. K., Leskinen, E., & Lyytinen, H. (2007). Reading development subtypes and their early characteristics. *Annals of Dyslexia, 57*(1), 3–32.
- Wiederholt, J. L., & Bryant, B. R. (2001). *Gray Oral Reading Tests: GORT-4*. Austin, TX: Proed.
- Woodcock, R. W. (1991). *Woodcock Language Proficiency Battery*. Itasca, IL: Riverside Publishing.

## Chapter 12

# Using Latent Change Score Analysis to Model Co-Development in Fluency Skills

Yaacov Petscher, Sharon Koon and Sarah Herrera

Measuring change over time is important in many contexts. It allows for a quantitative description of gains or losses that have occurred over time and, based on that evidence, future change under similar circumstances can be predicted. In an educational context, growth models, such as those described in Chap. 10 of this volume, can be used to measure individual growth over time and may utilize those expectations in future models to evaluate educational programs. Evaluation may proceed, for example, by comparing relative increases or decreases in a student's achievement to average or ambitious performance standards. Given the underlying importance of evaluating the development of an individual in terms of his or her educational and psychological outcomes, a plethora of models are readily available to researchers and practitioners; so many so, in fact, that McArdle (2009) concluded that all "repeated measures analyses should start with the question, 'What is your model for change?'" (p. 601). Answers to this question are predicated on several facets related to one's data, including the number of time points collected, the scale of measurement (e.g., ordinal or interval), and the distributional characteristics of the data. Having fewer time points restricts the classes of models to linear growth, while having more time points allows for the ability to model curvilinear trends with the caveat that the increased complexity of nonlinear models based on more time points often also requires more individuals. The scale of measurement for the data impacts the type of model that can be used to estimate one's developmental trajectory; ordinal data have more restrictive conditions for the identification of growth models than interval data (Mehta, Neale, & Flay, 2004; O'Connell, Logan, Pentimonti, & McCoach, 2013). Though

---

Y. Petscher (✉) · S. Koon · S. Herrera  
Florida Center for Reading Research, Florida State University, Tallahassee, FL, USA  
e-mail: ypetscher@fcrr.org

S. Koon  
e-mail: skoon@fcrr.org

S. Kershaw  
e-mail: sherrera@fcrr.org

© Springer Science+Business Media, LLC 2016  
K. D. Cummings, Y. Petscher (eds.), *The Fluency Construct*,  
DOI 10.1007/978-1-4939-2803-3\_12

ordinal data are less frequently observed with fluency outcomes, the number of time points and the nature of the distributions of scores are particularly germane to conversations about appropriate models for fluency change.

Fluency measures, such as curriculum-based measures, have established importance in educational research, especially as student growth is being studied as a potentially important variable for early identification of students who are at risk for later reading difficulties (e.g., Fletcher, Coulter, Reschly, & Vaughn, 2004). When attempting to model individual growth with fluency data, we typically have panel data that are not normally distributed. This characteristic of progress monitoring data may pose serious problems depending on your choice of model: (1) increased error rates in the identification of students using scores from universal screening assessments (i.e., students may be over- or under-identified), (2) issues with predictive validity as forecasting later reading disabilities become confounded by the presence of floor effects (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009), and (3) restrictions in evaluating individuals who are discrepant in both benchmark performance as well as growth.

Catts et al. (2009) found that floor effects were pervasive across measures of initial sound fluency, letter naming fluency, phoneme segmentation fluency, non-word fluency (NWF), and oral reading fluency (ORF). Moreover, it was observed that the predictive validity of the scores to later performance was heteroscedastic such that weaker associations were found between the criterion and predictor for students who were at the lowest end of the distribution. Though panel data were not evaluated in the study, it is not difficult to generalize the presence of floor or ceiling effects or, for that matter, broader non-normality (e.g., bimodality), to estimates of average growth. In this chapter we discuss a relatively new latent variable technique, latent change score (LCS) modeling (McArdle, 2009), which may be useful to researchers with panel data for fluency. Unlike traditional individual growth curve analysis, which estimates an average slope for the sample of individuals, LCS models segment time into multiple change scores which are then used for estimating average growth and causal effects. The goals of this chapter are to introduce the reader to the conceptual and mathematical underpinnings of LCS models, present information about how they differ from traditional latent growth models, and provide an illustration as to how they may be used to better understand individual differences in change over time using fluency data.

## Approaches to Growth Curve Modeling

Despite the fact that mean individual growth curves can be estimated from a variety of statistical models, differential interpretations will result based on the choice of model, which harkens back to McArdle's (2009) question.

When one has longitudinal data, two broad frameworks emerge as conventional methodologies for individual growth curve modeling, namely, multilevel regression and latent variable analysis. Many studies use one of these two approaches, yet as long as the same assumptions are met the models will yield identical results (Hox, 2000). This observation is due to the fact that the terminology of growth models (e.g., random effects growth models, hierarchical linear growth models, latent growth, mixed effects growth models) obfuscates the point that the model specification and estimation are often the same (Mehta & Neale, 2005; Branum-Martin, 2013). Consider the traditional multilevel model for growth where time is nested within the individual:

$$y_{it} = \pi_{0i} + \pi_{1i}(X_{it}) + e_{it}$$

$$\pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10} + r_{1i}$$

where  $y_{it}$  is the score for student  $i$  at time  $t$  on a measure  $y$ ,  $\pi_{0i}$  and  $\pi_{1i}$  characterize the initial status (based on centering) and slope, respectively, where  $\beta_{00}$  and  $\beta_{10}$  are the means corresponding to the status and slope,  $X_{it}$  is a variable that describes time, and is coded in a way to reflect the measurement occasions (e.g., 0, 1, 2 for three time points centered at time 1),  $r_{0i}$  and  $r_{1i}$  are the random effects for the two parameters, and  $e_{it}$  is the measurement-level residual. Consider as well the specification of the same growth model in a latent framework:

$$y_{it} = \lambda_{0t}\eta_{0i} + \lambda_{1t}\eta_{1i} + \varepsilon_{it}$$

$$\eta_{0i} = \nu_0 + \zeta_{0i}$$

$$\eta_{1i} = \nu_1 + \zeta_{1i}$$

The outcomes from both specifications are the same, and the structures of the equations are nearly identical. Rather than initial status, slope, and random effects being represented by the  $\pi$ ,  $\beta$ , and  $r$  components, as in the multilevel regression, in the latent framework they are characterized with  $\eta$ ,  $\nu$ , and  $\zeta$ , respectively. The covariate  $X_{it}$  in the multilevel regression, which denoted the coding of measurement occasion, is replaced by  $\lambda_{1t}$ , which represents factor loadings coded in the exact same way (e.g., 0, 1, 2 for three time points centered at time 1).  $\lambda_{0t}$  in the model represents factor loadings for the intercept, and is constrained to values of 1 for each of the time points in a model. Several authors have demonstrated that identical results may be obtained using either the multilevel regression or latent growth curve approach (Stoel, van Den Wittenboer, & Hox, 2004).

As a full measurement model, consider the latent growth model in Fig. 12.1, which includes latent intercept (i.e., I) and slope (i.e., S) factors, which are indicated by four measurement occasions (T1–T4) each with a residual error term ( $\epsilon_i$ ). Both of the latent factors have a mean ( $\eta_{0i}$  and  $\eta_{1i}$ ), a variance ( $\psi_{0i}$ ,  $\psi_{1i}$ ), and a covariance between the two ( $\psi_{01i}$ ). This type of model may answer the same types of questions as a multilevel growth model, such as:

- What is the average growth rate for a group of individuals?
- To what extent do individuals vary in their rate of change?
- What predictors explain variance in growth rate for a group of individuals?
- What is the relation between how students perform at the beginning of data collection and growth over the time period of data collection?

Although such questions are readily addressed by either modeling approach, there are limitations to traditional multilevel regression as implemented in many conventional software packages. Namely, questions concerning structural causality or multivariate longitudinal analysis cannot be directly modeled in a multilevel regression framework. The latent approach relaxes such constraints to allow for such questions to be addressed because it is able to model multivariate outcomes, whereas the multilevel framework is traditionally restricted to univariate outcomes (Muthén, 2004). In addition, latent variable modeling allows the constructs to be designated as either continuous or categorical factors, allowing for the analysis of different classes.

The flexibility of the latent framework for univariate or multivariate repeated measures data substantially increases the types of models that can be estimated,

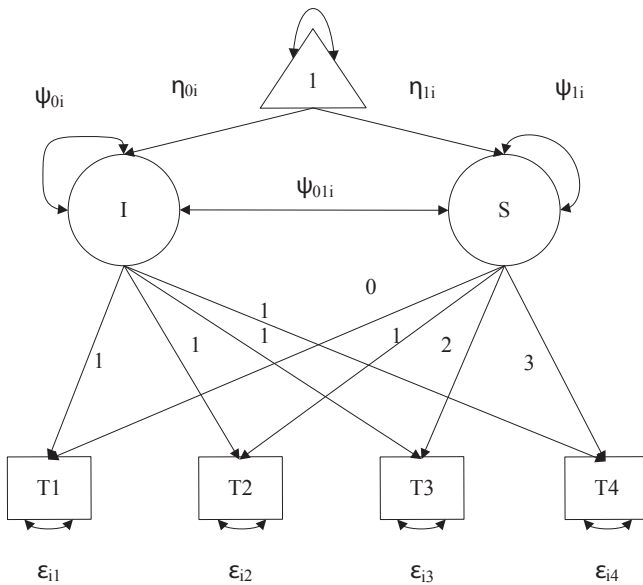


Fig. 12.1 Sample linear latent growth curve model

including those that combine aspects of factor analysis, time series, and multivariate analysis of variance. Beyond the questions of individual growth and predictors of individual differences in growth, the latent variable framework allows for the investigation of the determinants (i.e., causes) of individual change, as well as the determinants of individual differences in individual change (McArdle & Grimm, 2010). Individual growth curve analyses are not equipped to answer these questions as they are fitting individual curves to the repeated measures data.

A number of classic models have been used to address questions of causality in longitudinal data. One of the earliest forms of testing effects of longitudinal data is the autoregressive (or simplex) model (Joreskog, 1970; Joreskog & Sorbom, 1979). This model tests a chain of multiple assessment periods, where each time point is regressed on the immediate preceding assessment. The inherent design of the simplex model is such that the correlation between time  $t$  and  $t-n$  (where  $n$  is a preceding measurement occasion) is expected to decrease as the distance between  $t$  and  $n$  increases. This expectation dictates that the correlation between the first and last time point will be the weakest association in the series of measurement occasions. Subsequently, the autocorrelation measures the stability of change between individuals; a coefficient near zero indicates poor stability between the measurement occasions, whereas larger coefficients indicate that individuals' have maintained their rank ordering over time (Bollen & Curran, 2004). The simplex model may be specified as a univariate or multivariate model. An extension of the multivariate autoregressive model is the cross-lagged model (Gollob & Reichardt, 1987). The cross-lags represent the unique longitudinal effect of one variable on the other after controlling for the autoregressive effect. In general, the autoregressive model is a flexible framework for estimating structural causality among measures over time. It can easily handle observed measures of latent variables as well as univariate or multivariate outcomes. Although the autoregressive model is able to estimate the regression of time points on each other, it is unable to fit individual growth curves for the longitudinal data (Hertzog & Nesselrode, 1987).

## Model Selection

By modeling longitudinal data in a latent framework, there is much greater flexibility to estimate individual trajectories, understand determinants of change, and test the extent to which multivariate outcomes reciprocally relate to change. Once a decision has been made to utilize a latent variable framework, as opposed to traditional multilevel regression, several decisions must be made concerning the type of latent variable to fit to one's data. One must initially evaluate whether the primary question of interest is concerned with testing causality with the panel data, or in estimating average change with the panel data, because latent variable models have typically restricted data analyses to be either of the causal nature or the type which



estimates average change. This gives pause as to which class of latent variable models is appropriate to answer the question, given the mutually exclusive nature of model specification and estimation.

As an example, suppose an individual has collected monthly progress monitoring data on ORF over the course of a school year. It is possible that the question of interest would relate to the effects of a previous assessment period on the next (e.g., October ORF on November ORF), in which case the simplex model may be best suited to the analysis. Though it would not be possible to estimate the average rate of growth across participants, an inherent question embedded in the autoregressive model is that of stability from one measurement occasion to the next. Thus, while the researcher may obtain multiple estimates corresponding to the stability of performance from one occasion to the next, such estimates may not be as reliable as when leveraging all of the measurement occasions to construct the individual growth curve. Similarly, if the researcher also collected progress monitoring data on a measure of mathematics computation fluency (MCF), the simplex model of ORF could be extended to include MCF scores. This approach will still allow for an evaluation of the univariate effects of the prior fluency score on the next time point, but could additionally include cross-lags to estimate the effect of October ORF on November MCF, and October MCF on November ORF.

Conversely, should the question of interest be centered on characterizing individual growth, individual growth curves may be used to model average growth and variance in a univariate dimension for just ORF, a multivariate latent growth curve to model simultaneous growth in ORF and MCF, as well as multivariate growth curve models with specific regressions to estimate the effect of ORF initial status on MCF growth, as well as MCF initial status on ORF growth. In these instances the researcher is able to characterize trends in fluency, but is unable to evaluate causality within each measure.

## **The Latent Change Score (LCS) Model**

Given the noted limitations of both traditional structural analyses as well as longitudinal models for panel data, the LCS model (McArdle, 2009; McArdle & Hamagami, 2001; McArdle & Nesselroade, 1994) was developed to combine causal and individual growth curve analysis. While traditional latent growth curve and multi-level regression analyses are useful in yielding average estimates of change over time, they are limited in that causal factors cannot be modeled pertaining to growth, nor can growth be segmented into piecewise “chunks” of change to evaluate unique effects of average growth or change. It is certainly plausible in the field of CBM research to observe fluency data that demonstrate differential change across extended progress monitoring assessment. When students differentially change, whether due to immediate intensive interventions, measurement sensitivity to skills development, or individual student factors, average estimates of growth may be insufficient for characterizing the nature of the data. The LCS model allows one to estimate not

only the average growth but also the average change scores across the assessment periods. In this way it possible to evaluate whether students differentially change between any two segments of assessments within a wider data collection period.

The LCS model represents a flexible approach that simultaneously estimates additive change over time (i.e., the average growth across change scores;  $\alpha$ ) and multiplicative change (i.e., proportional change;  $\beta$ ). Before providing model equations and explication of the underlying components to the model, it is first useful to picture the representation of the LCS model and understand the path and latent components. A general path diagram of a *univariate* LCS model is given in Fig. 12.2. Similar to Fig. 12.1, there are latent factors for intercept and slope, observed measures with unique effects for the four time points, as well as means, variances and a covariance for the latent intercept and slope factors. The unique components of the LCS model include latent factors for the observed measures at each time point, autoregressive effects for the time-specific latent factors, LCS factors ( $\Delta T12$ – $\Delta T34$ ) with associated loadings (i.e.,  $\alpha$ ), and proportional change effects (i.e.,  $\beta$ ). An underlying mechanism of the LCS model lies in its origins in classical test theory, namely, that one's observed score at a time point ( $Y_{it}$ ) is modeled as a function of an unknown latent true score ( $y_{it}$ ) and a unique score for the individual ( $e_{it}$ ), and is expressed as:

$$Y_{it} = y_{it} + e_{it}.$$

In Fig. 12.2,  $y_{it}$  is expressed via the observed measures of T1–T4,  $Y_{it}$  is represented by the latent factors T1–T4, and  $e_{it}$  are the errors  $\varepsilon_{i1} - \varepsilon_{i4}$ . As it pertains to the LCS, recall that the simple difference score for an observed measure ( $\Delta Y_i$ ) is calculated as the difference between a specific time point ( $Y_{it}$ ) and performance at an earlier time point ( $Y_{(t-1)i}$ ):

$$\Delta Y_{it} = Y_{it} - Y_{(t-1)i}.$$

This equation can be extended to a latent variable model as

$$\Delta y_{it} = y_{it} - y_{(t-1)i} \quad (\text{Eq. 12.1})$$

or could be further rearranged to be expressed as

$$y_{it} = y_{(t-1)i} + \Delta y_{it} \quad (\text{Eq. 12.2})$$

which expresses that a latent score  $y$  for individual  $i$  at time  $t$  (i.e.,  $y_{it}$ ) is comprised of a latent score from a previous time point ( $y_{(t-1)i}$ ), and the amount of change which occurs between the two points ( $\Delta y_{it}$ ). When considering the first two time points in Fig. 12.2, the autoregressive effect of T1 on T2 is fixed at 1, which yields a difference score estimation in Eq. 12.1 that is more simplistic in its estimation (i.e., uses simple subtraction; McArdle 2009). The LCS of  $\Delta T12$  is not directly measured (unlike T1 and T2), and may be characterized as the portion of the T2 score which

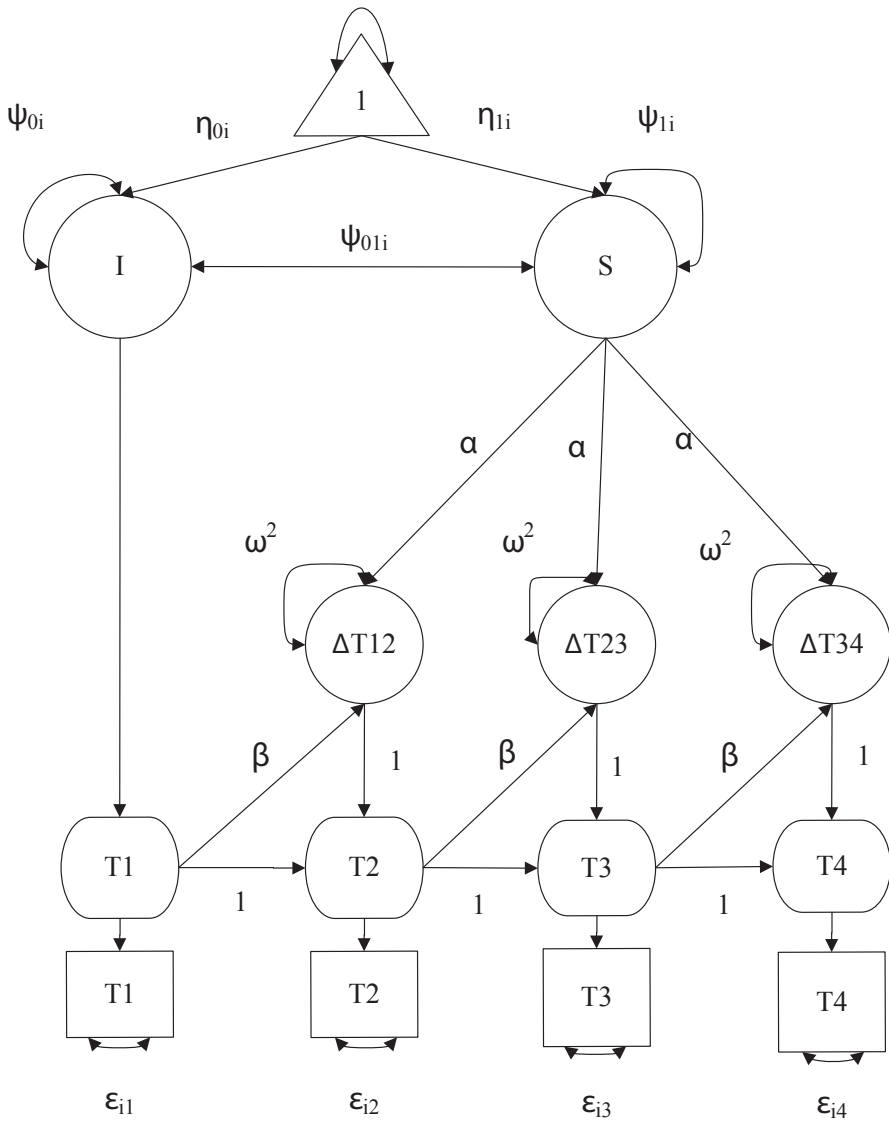


Fig. 12.2 Sample univariate latent change score (LCS) model

is not equal to  $T_1$  (McArdle & Nesselroade, 1994). A second constraint related to the change score involves the regression of  $T_2$  on the change score (i.e.,  $\Delta T_{12}$ ) being fixed at 1. This constraint serves to identify the model as well as allow for the estimation of the LCS mean and variance.

Equation 12.1 serves to illustrate how the individual LCSs are estimated, and from these allow for the growth and causal effects to be estimated. The growth portion of the LCS model is based on the change score loadings (usually fixed to 1)

associated with latent slope factor (i.e.,  $\alpha$ ). Note that, different from Fig. 12.1, the slope is not indicated by the observed measures but rather the LCSs. The growth portion of the LCS model is then estimated as a function of the multiple LCSs:

$$y_{it} = \eta_{0i} + \Sigma(\Delta y_{it}) \quad (\text{Eq. 12.3})$$

Note the difference between Eqs. 12.2 and 12.3 in that Eq. 12.3 includes an estimated intercept term,  $\eta_{0i}$ , which denotes the individuals' initial status based on centering (much like  $\eta_{0i}$  in the latent growth model), and  $\Sigma(\Delta y_{it})$  represents the summed LCSs up to time  $t$  (Grimm, 2012). The three LCSs load on the slope factor with  $\alpha$ , which are typically fixed at 1 for model estimation. The causal portion of the LCS model is specified by the regression of the change score on the previous time point,  $\beta$ , which is also known as the autopropotion effect, and a residual,  $\omega^2$ . When both  $\alpha$  and  $\beta$  are estimated in the LCS model, it is referred to as a *dual change score* model; dual in the sense that both constant change (i.e., average change) via the  $\alpha$  coefficients, and proportional change via the  $\beta$  coefficients are simultaneously estimated. When the dual change score model is estimated, Eq. 12.1 is extended to:

$$\Delta y_{it} = \alpha * \eta_{it} + \beta * y_{(t-1)i}, \quad (\text{Eq. 12.4})$$

which essentially states that the LCS  $\Delta y$  at time  $t$  for individual  $i$  is estimated as a function of the average latent change slope ( $\alpha * \eta_{it}$ , where  $\alpha=1$ ), plus the proportional change ( $\beta * y_{(t-1)i}$ ). The dual change score model represents the most complex specification of an LCS model for a univariate outcome; however, the  $\alpha$  and  $\beta$  parameters may be differentially restricted given one's research questions. If a primary goal is to solely estimate constant change, the  $\beta$  coefficients can be fixed to 0. This reduces the dual change model from Eq. 12.4 to what is termed a *constant change model* and is estimated with:

$$\Delta y_{it} = \alpha * \eta_{it}.$$

Conversely, when only the proportional change component of the dual change score model is of interest, the  $\alpha$  coefficients are fixed to 0, as is the mean and variance of the latent slope, thus reducing Eq. 12.4 to:

$$\Delta y_{it} = \beta * y_{(t-1)i}.$$

The univariate specification of the LCS model can be extended to handle multiple outcomes. While one measure of fluency, such as NWF, could be modeled in the univariate LCS framework, individual data on non-word and ORF could be fit in a bivariate LCS model. Figure 12.3 illustrates a general model for a *bivariate dual change score model*, where the primary difference between it and the univariate specification is the inclusion of what are called *coupling* or *cross-lag effects* (i.e.,  $\gamma_{yx}$  and  $\gamma_{xy}$ ). The insertion of the coupling effect in the bivariate LCS equations extends Eq. 12.4 to

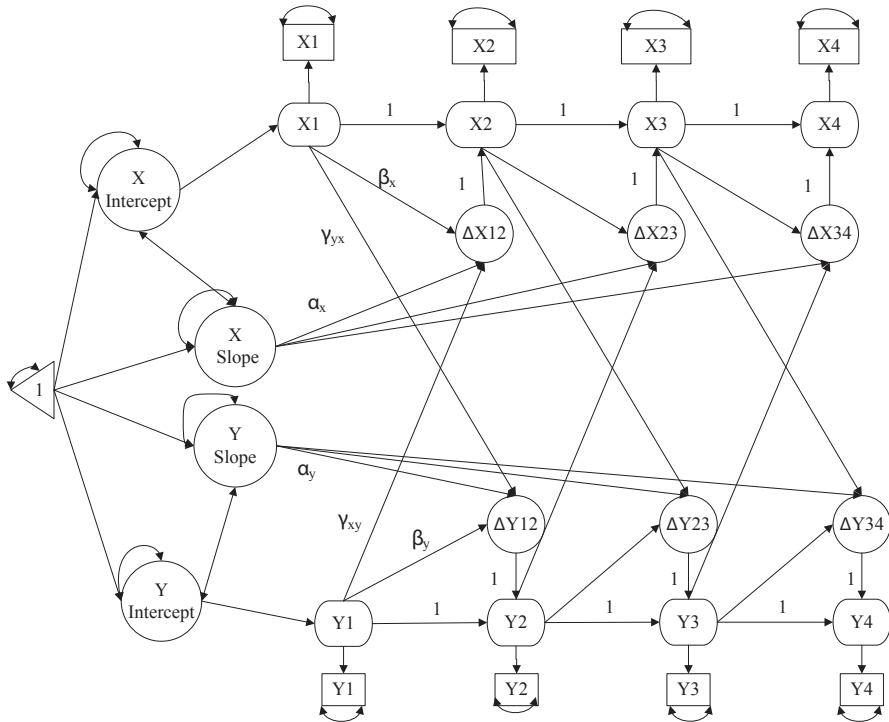


Fig. 12.3 Sample bivariate dual change score model

$$\begin{aligned} \Delta y_{ti} &= \alpha_y * \eta_{1,yi} + \beta_y * y_{(t-1)i} + \gamma_{yx} x_{(t-1)i} \quad \text{and} \\ \Delta x_{ti} &= \alpha_x * \eta_{1,xi} + \beta_x * x_{(t-1)i} + \gamma_{xy} y_{(t-1)i}. \end{aligned} \tag{Eq. 12.5}$$

As with the univariate LCS model in Eq. 12.4, Eq. 12.5 includes subscripts for the constant and proportional change coefficients that are specific to the outcome for which the LCS is estimated. Further, the insertion of the coupling effect (e.g.,  $\gamma_{yx} x_{(t-1)i}$ ) denotes the effect of  $x$  from a previous time point (e.g., X1 in Fig. 12.3) on the change score for the current time point (e.g.,  $\Delta Y12$ ). The flexibility of the bivariate model is such that it can be expanded to multivariate scenarios that are only hindered by sample size and computing power. Similar to the univariate model, the bivariate LCS model contains flexibility to allow the constant change, proportional change, or coupling effects to be freed or fixed if the question of interest does not require all three components to be freed for estimation.

### LCS Model Development and Evaluation

Development of the LCS model is flexible so that one may specify a fully unconstrained dual change model and through a set of constraints on the  $\alpha$ ,  $\beta$ , and  $\gamma$

parameters can test the extent to which a constant change, proportional change, or dual change score model results in the most parsimonious model applied to the data. Evaluation of model parsimony for the LCS models is no different than conventional methods used for latent variable analyses. Criterion-based fit indices such as the comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), standardized root mean residual (SRMR), as well as relative fit indices such as the AIC and BIC are appropriate for model specification.

### *How the LCS Model May Inform Fluency Research*

To illustrate the importance of LCS models in studying reading fluency, one can look at published research on reading fluency using latent growth curve analysis or multilevel regression models and envision how applying an LCS framework could further lead to comprehensive understanding of the data and, as a result, increase the usefulness of stated conclusions based on the data. Al Otaiba, Petscher, Pappamihiel, Williams, Drylund, & Connor (2009) estimated Latino students' ORF growth trajectories over a 2-year period using a nonlinear, piecewise growth curve model. Each piece of the model estimated a growth trajectory for ORF within 1 school year. Differences in average growth rates were related to English proficiency levels as well as special education status. While knowledge of differential average growth rates across school years could be used to develop appropriate grade level benchmarks for use in screening students for additional intervention, an LCS model could better inform the proportion of growth occurring at each time point within each school year to provide even more specific targets and timely intervention services.

Similarly, Kim, Petscher, Schatschneider, & Foorman (2010) estimated initial status and growth rates in phonological decoding skills and ORF for students in grades 1–3, using a linear piecewise growth curve model. Relations between initial status and growth rates were explored and the results revealed differential growth rates at the student level. Specifically, students in grade 1 with high ORF initial status were more likely to grow faster than those with a lower fluency rate, evidenced by a large positive correlation between the initial status and the growth rate. This relation changed in grades 2 and 3, with a negative correlation indicating that students with high ORF initial status tended to have slower growth over the year. Moreover, the authors found that the individual growth estimate was uniquely predictive of performance on the reading comprehension subtest of the Stanford Achievement Test-10th edition above that predicted by a status measure (i.e., ORF performance at the fall). A conclusion drawn by the authors was that growth may be an important factor to consider in understanding individual differences in reading comprehension performance in grades 1–3.

Several advantages of using the LCS model may exist when applied to the Kim et al. (2010) study. As an LCS (Eq. 12.4) views development over time as a function of average change plus proportion change, it is possible that two LCSs estimated from a growth trajectory may differentially explain individual differences in an outcome. That is, a linear individual growth curve is estimated as a function of,

minimally, three waves of panel data. These time points could instead be chunked as two difference scores (i.e., T2–T1 and T3–T2). Subsequently, if the magnitude of change varies between the two difference scores, one may explain greater variability in outcomes than the other. In this way, an LCS model may not only be useful for better understanding student differences in growth such as in the Al Otaiba et al. (2009) study, but may also be useful in more comprehensively evaluating individual differences in proximal and distal outcomes.

An LCS approach also could inform differential ORF growth trends in the context of evaluating reading instructional materials. This context was the focus of a study conducted by Crowe, Connor, & Petscher (2009) for the purpose of investigating the effect of six core reading curricula on students' ORF growth and whether this effect varied by students' grade or SES status. Differences in average growth rates were found across curricula, as well as between groups of students with different characteristics. Expanding this analysis to an LCS framework would allow for identification of curricular materials that are most effective within specific time frames. Additional examples could be provided that would further showcase the importance of LCS models in studying reading fluency growth rates in particular and educational issues in general. However, the above examples are likely sufficient to demonstrate that the LCS framework is useful when there is a possibility that changes occur differentially over time and knowledge of the differential rates is important to understanding the processes or conditions leading to the change.

## Applied Example

To build a better understanding of how these models may be developed and evaluated to describe change over time the latent growth, constant change, proportional change, and dual change score models will be illustrated in the following sections. For all modeling applications, criterion fit indices (i.e., CFI, TLI, RMSEA) are reported to evaluate the extent to which the specified model provides parsimonious fit to the data according to conventional thresholds. The CFI and TLI values of at least .95 are considered acceptable, and RMSEA up to .10 also provide evidence of acceptable model fit. Between-model comparisons were made by using either a  $\chi^2$  difference test when models were nested or using the BIC to evaluate model parsimony when models were non-nested. Raftery (1995) demonstrated that between-model BIC differences between 10 and 100 are sufficient to indicate a practically important difference in model fit.

A few notes on our approach to modeling, and the explication of model comparisons bear a mention. First, conventional approaches to latent growth curve modeling assume that the residual variances of the observed variables are the same over time. This is known as the assumption of homoscedasticity. More recent evaluation of this assumption has suggested that such a constraint does not contribute much to the understanding of important model estimates including latent factor means and variances, yet it does contribute to substantial misfit and estimation of



the variances and covariances of the model (see Grimm & Widaman, 2010; Sivo, Fan, & Witta, 2005).

Second, research which has used LCS models (e.g., Malone et al., 2004; Reynolds & Turek, 2012) has typically constrained the autoproportion effects to be equal over time; this is done from a theoretical standpoint so that one may assume that the dynamic relation does not change over the observed developmental period. Though the relation between prior fluency performance and change over time might be expected to be static, it could also be viewed as an empirical question. Thus, in the context of the present application, four dual change score models were estimated with: (1) constrained error variances and autoproportions; (2) freed error variances and constrained autoproportions; (3) constrained error variance and freed autoproportions; and (4) freed error variances and autoproportions.

## *Participants*

Data for the following set of examples were obtained from the Progress Monitoring and Reporting Network at the Florida Center for Reading Research. Students in this database were assessed between three and four times a year on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good, Kaminski, Smith, Laimon, & Dill, 2001) assessments as part of Florida's assessment system during the federal *Reading First* initiative which occurred from 2003 to 2009 (Foorman, Petscher, Lefsky, & Toste, 2010). The present data comprised 73,916 second-grade students who had progress monitoring data on the DIBELS NWF and ORF assessments. These second-grade students were administered the NWF and ORF assessments four times during the 2005–2006 academic year during the months of September, December, February, and April.

## *Measures*

The DIBELS NWF (Good et al., 2001) is a standardized assessment designed to measure an individual's ability to blend letter sounds into words. The student is presented with VC and CVC nonsense words that are randomly ordered and is asked to either pronounce each sound individually or to say the whole word. The student is given 1 min to state as many of the presented sounds as possible and the total score is based on the number of correct letter sounds produced within that time frame. Alternative form reliability is strong (.83) and is strongly correlated with DIBELS ORF ( $r = .82$ ; Cummings, Dewey, Latimer, & Good, 2011)

The DIBELS ORF (Good et al., 2001) is a measure that assesses oral reading rate and accuracy in grade-level connected text. This standardized, individually administered test was designed to identify students who may need additional instructional support in reading and to monitor progress toward instructional goals (Good & Kaminski, 2002). During a given administration of ORF, students are asked to read aloud

three previously unseen passages consecutively, for 1 min per passage. Students are given the prompt to “be sure to do your best reading” (Good et al., 2001, p. 30). Between the administration of each passage, students are given a break, in which the assessor simply reads the directions again before the task resumes. Words omitted, substituted, and hesitations of more than 3 s are scored as errors, although errors that are self-corrected within 3 s are scored as correct. Errors are noted by the assessor, and the score produced is the number of words correctly read per minute (wcpm). From the three passages, the median score is used for decision-making about level of risk and level of intervention needed. Information about how the risk levels for ORF benchmarks were developed and what ranges of scores correspond to various levels of risk are available from several technical reports by the DIBELS authors (e.g., Good, Wallin, Simmons, Kameenui, & Kaminski, 2002). Speece and Case (2001) reported parallel form reliability of .94, and research has demonstrated adequate to strong predictive validity of DIBELS ORF for reading comprehension outcomes ( $r = .65-.80$ ; Petscher & Kim, 2011; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008).

## *Software*

There are a number of software packages available which can appropriately estimate LCS models including LISREL (Joreskog & Van Thillo, 1972), AMOS (Arbuckle, 2006), and Mplus (Muthén & Muthén, 1998–2012), as well as the RAMpath package (Zhiyong, McArdle, Hamagami, & Grimm, 2013) in R. For the present illustrations, all statistical models were run in Mplus and figures were generated in R. Data scripts for each of the generated models are available from the first author.

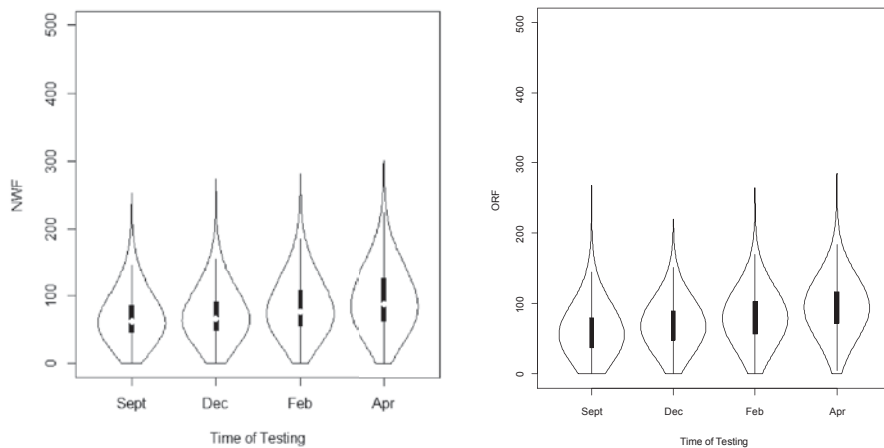
## *Results*

The goal of this illustration is to highlight the LCS model as well as its comparison to traditional latent growth curve analysis. Univariate latent growth curve models will first be applied to the NWF and ORF data in order to evaluate both fit as well as the average rates of growth. Next, a proportional change LCS model will be fit, followed by a constant change model, a dual change score model, and finally a bivariate dual change model. Given the large sample included in this example, we randomly selected ten students in order to highlight differences in the observed and estimated trends across the models.

**Descriptive Statistics** Sample statistics for the NWF and ORF measures across the four time points are provided in Table 12.1. Both measures demonstrated relatively normal score distributions, yet some skew and kurtosis existed for the NWF measures. Graphing the descriptive statistics as violin plots (Fig. 12.4) better displays the statistical summary from Table 12.1, where it may be observed that NWF was more likely to include scores further from the mean compared to ORF. Violin plots are a useful mechanism for simultaneously evaluating the distribution of scores and

**Table 12.1** Descriptive summary for non-word fluency (NWF) and oral reading fluency (ORF) at each of the four assessment periods

Measure	Min	Max	Mean	S.D.	Skew	Kurtosis
NWF fall	0	253	65.95	31.66	0.99	1.11
NWF winter 1	0	274	80.92	37.20	0.95	1.10
NWF winter 2	0	281	87.09	38.62	0.79	0.57
NWF spring	0	300	95.35	42.37	0.66	0.29
ORF fall	0	220	56.09	31.67	0.63	0.44
ORF winter 1	0	220	69.18	31.82	0.37	0.21
ORF winter 2	0	261	81.70	34.58	0.22	0.16
ORF spring	0	248	92.70	35.25	0.06	0.28



**Fig. 12.4** Violin plot for NWF and ORF raw scores

the interquartile range of scores. The plots for both measures highlight that average performances increased across the four testing periods; however, it appeared that the interquartile range for NWF increased across the year whereas it remained relatively stable for ORF.

**Growth Models** Prior to fitting the growth models, it was of interest to plot the raw data to evaluate whether the scores for each of the measures demonstrated a linear or curvilinear trend. Ten students were randomly selected and their plotted scores for NWF and ORF are presented in Fig. 12.5. Both within and across the measures it is clear that individual differences in trends exist. Although some of the randomly selected students demonstrated a linear growth pattern, most individuals’ growth could be characterized as nonlinear (see Chap. 10, this volume). When the individual scores are grouped together in one plot for each assessment (Fig. 12.6), the differences in individual observed performances is more readily evaluated. Individual differences across the time points were larger for NWF compared to ORF, with several students demonstrating a drop off in performance between the third

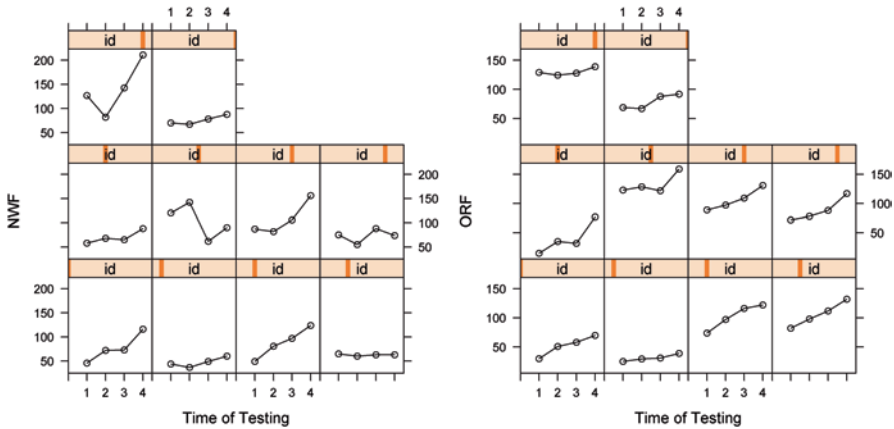


Fig. 12.5 Individual raw score plots of non-word fluency (*NWF*) and oral reading fluency (*ORF*)

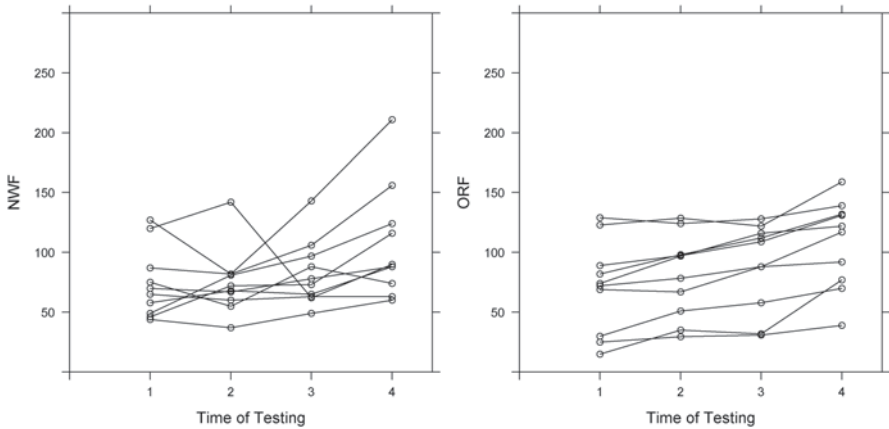


Fig. 12.6 Grouped raw score plots of non-word fluency (*NWF*) and oral reading fluency (*ORF*)

and fourth time points. Such differences in the sample plots suggested it may be valuable to test a curvilinear growth model in addition to the linear latent growth and LCS models.

### Latent Growth Models

In Table 12.2, results are presented for the fit of the linear latent growth model for *NWF* and *ORF*. Across all of the indices, the linear model provided adequate fit to the data; however, the RMSEA and associated confidence interval was outside of the typically accepted upper bound (i.e., .10). Figure 12.7 presents the model coefficients for each of the respective outcomes. Based on the centering of the models at

**Table 12.2** Fit indices for model testing

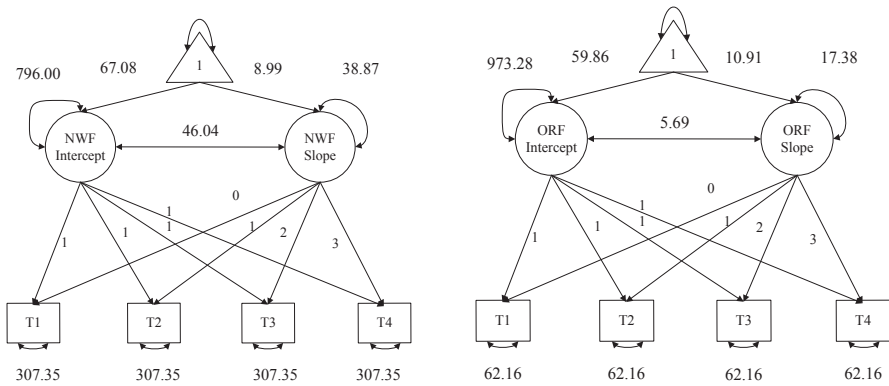
Outcome	Model	$\chi^2$	df	CFI	TLI	BIC	RMSEA	LB	UB	$\Delta\chi^2$	$\Delta df$	p
NWF	Linear	7,822	8	0.96	0.97	2,757,813	0.115	0.113	0.117			
	Nonlinear	2,911	4	0.99	0.98	2,752,935	0.099	0.096	0.102	4,910	4	<.001
	Constant change	7,822	8	0.96	0.97	2,757,815	0.115	0.113	0.117			
	Proportional change	10,122	10	0.95	0.97	2,760,099	0.117	0.115	0.119	2,300	2	<.001
	Dual change 1	4,233	7	0.98	0.98	2,754,233	0.090	0.088	0.093	3,589 <sup>a</sup>	1	<.001
	Dual change 2	2,257	4	0.99	0.99	2,752,282	0.087	0.084	0.090	1,976 <sup>b</sup>	3	<.001
	Dual change 3	2,423	5	0.99	0.99	2,752,439	0.081	0.078	0.084	1,810 <sup>b</sup>	2	<.001
	Dual change 4	523	2	0.99	0.99	2,750,565	0.059	0.055	0.064	1,734 <sup>c</sup>	2	<.001
	Linear	42,822	8	0.92	0.93	2,433,388	0.269	0.267	0.271			
	Nonlinear	30,334	4	0.94	0.91	2,420,932	0.320	0.317	0.323	12,488	4	<.001
ORF	Constant change	42,822	8	0.92	0.94	2,433,388	0.269	0.267	0.271	54,471	4	<.001
	Proportional change	97,293	10	0.81	0.89	2,487,843	0.363	0.361	0.365			
	Dual change 1	34,947	7	0.93	0.94	2,425,520	0.260	0.258	0.262	7,875 <sup>a</sup>	1	<.001
	Dual change 2	2,465	4	0.99	0.99	2,393,063	0.091	0.088	0.094	32,482 <sup>b</sup>	3	<.001
	Dual change 3	37,388	5	0.93	0.91	2,427,978	0.318	0.315	0.321			
	Dual change 4	4,668	2	0.99	0.97	2,395,281	0.178	0.173	0.182	30,279 <sup>b</sup>	5	<.001

Dual Change 1 = Constrained error variances and autoproporition, Dual Change 2 = Freed error variances and constrained autoproporitions, Dual Change 3 = Constrained error variances and freed autoproporition, Dual Change 4 = Freed error variances and autoproporition

<sup>a</sup> comparison made to constant change model

<sup>b</sup> comparison made to dual change 1

<sup>c</sup> comparison made to dual change 2

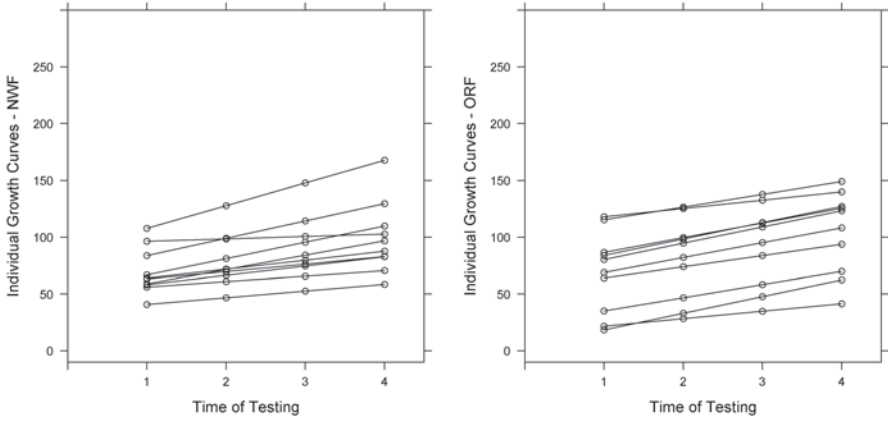


**Fig. 12.7** Linear latent growth curve models for non-word fluency (*NWF*) and oral reading fluency (*ORF*)

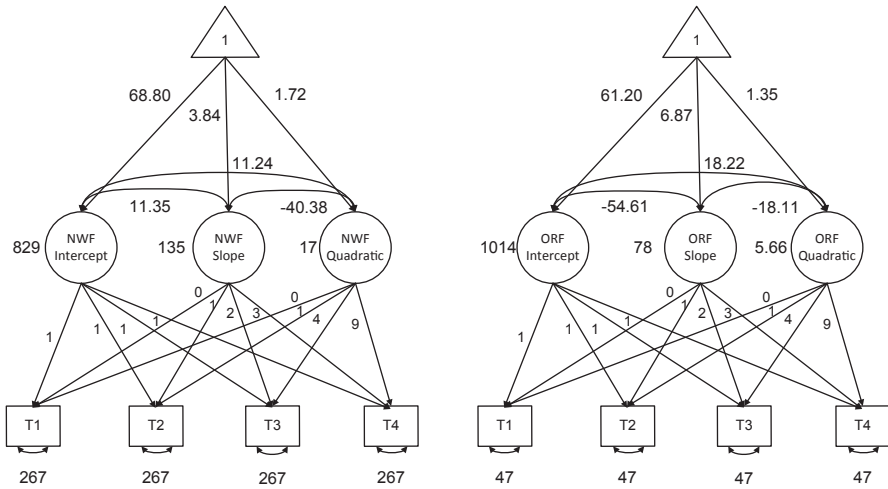
time 1 (i.e., September), the average NWF performance at the beginning of second grade was 67 correct letter sounds per minute (cls) and the average rate of growth was 8.99 cls between each assessment period.

Both NWF intercept and slope factors demonstrated significant variance (796.00 and 38.87, respectively), though considerably more was due to individual differences at the beginning of the year (95%). The variances for both factors were statistically significant indicating that individuals varied in both their September fluency rates as well as in their rate of growth. Given the raw score plots in Figs. 12.5 and 12.6, this was not surprising. The positive covariance between intercept and slope factors indicated that students who began second grade with higher NWF scores grew more than individuals with lower NWF scores; however, it is often difficult to gauge the relative magnitude of the effect. When converted to a correlation, the relation between the two was estimated at  $r = .26$ , or a moderate association between a student’s initial status and their rate of growth.

The linear latent growth model for ORF largely mirrored that of NWF. Average ORF scores of 60 were estimated for the sample with an average growth rate of 17.38 wcpm per assessment period. Significant variance was observed in both intercept and slope factors, with 98% of the total variance due to individual differences at the beginning of the year. When the covariance between intercept and slope was calculated as a correlation ( $r = .04$ ), it can be observed that there was no relation between initial status and slope. Data from the intercept and slope factors may be used to construct individual growth trajectories such as those presented in Fig. 12.8. When compared to the plotted observed data in Fig. 12.6, it can be seen that the estimated individual growth curves are not congruent to the raw data. Such a phenomenon is certainly expected to occur when modeling observed versus predicted trajectories using a linear approach, because the linear growth model constrains growth to be linear, and any inherent nonlinear trend, such as in the observed data plots will not manifest the more restricted linear growth model.



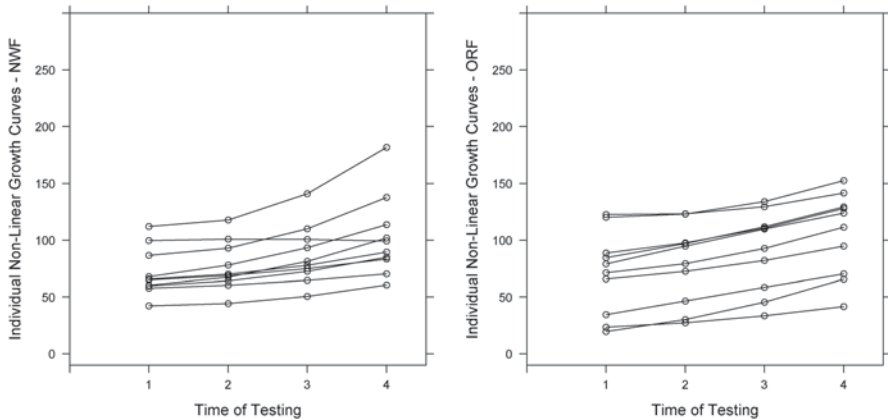
**Fig. 12.8** Grouped linear latent growth curve plots of non-word fluency (*NWF*) and oral reading fluency (*ORF*)



**Fig. 12.9** Nonlinear latent growth curve models for non-word fluency (*NWF*) and oral reading fluency (*ORF*)

Because four time points were available, a nonlinear latent growth model allowed for a more sophisticated evaluation of change in each of the fluency measures. As illustrated in Fig. 12.9 and noted in Chap. 10 of this volume, as well as in sources elsewhere (O’Connell & McCoach, 2008; Bowles & Montroy, 2013), the latent growth model may be extended to a basic nonlinear growth model by adding a third latent factor (i.e., the quadratic factor) to estimate the curvilinearity in the data. Note that the factor loadings for the quadratic factor are simply the square of the factor loadings from the slope factor.

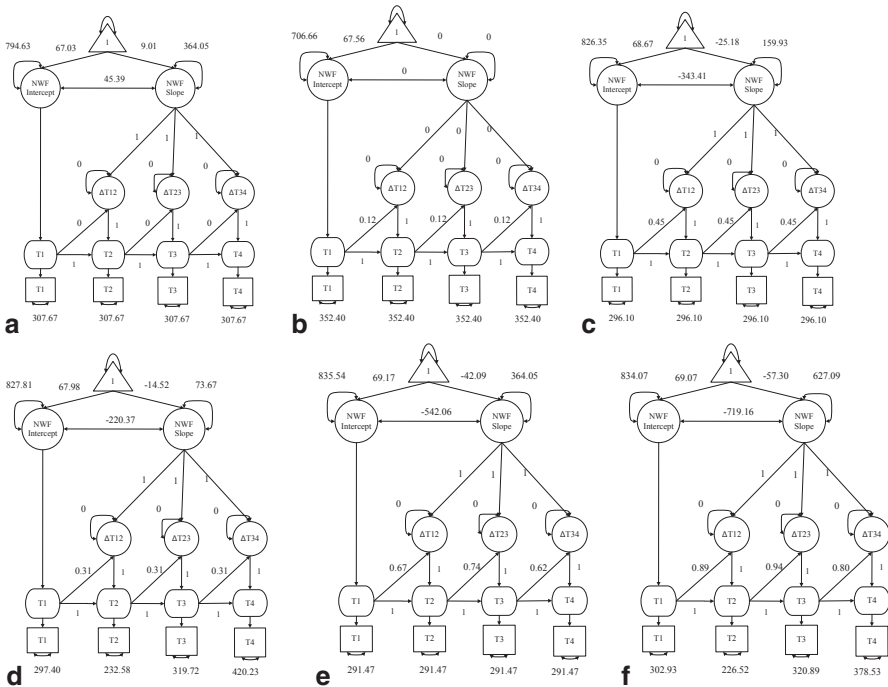




**Fig. 12.10** Grouped nonlinear latent growth curve plots of non-word fluency (*NWF*) and oral reading fluency (*ORF*)

The fit of the nonlinear model was superior to that of the linear model (Table 12.2). Note that the reduction of the  $\chi^2$  from the linear to the nonlinear model was statistically significant for both NWF ( $\Delta\chi^2=4,910$ ,  $\Delta df=4$ ,  $p < .001$ ) and ORF [ $\Delta\chi^2=12,488$ ,  $\Delta df=4$ ,  $p < .001$ ]. Coefficients for the nonlinear latent growth models are provided in Fig. 12.9. The intercepts for both NWF and ORF are approximately the same as when they were estimated in the linear model, with differences attributed to the addition of another factor. The primary difference between the linear and nonlinear models may be seen in the estimated means of the slopes; in the NWF linear model the average rate of growth was 8.99 cls per assessment period compared with 3.84 cls in the nonlinear model. This difference is accounted for by the inclusion of the quadratic factor; thus, in the nonlinear model, both the slope and quadratic factors are required to understand growth rates in the fluency scores. Pertaining to NWF, the average linear rate of growth was 3.84 cls with a quadratic rate of change of 1.72. These two pieces of growth information indicated that not only did students grow positively (via the linear rate of change), but they also accelerated their rate of NWF over the course of the school year (via the quadratic rate of change). Similarly, the growth rate for ORF was positive at 6.87 wcpm and accelerated at a rate of 1.35 wcpm per assessment period over the year. The effect of the nonlinear term on the individual growth curves can be seen in Fig. 12.10. When contrasting the plots with Figs. 12.6 and 12.8, it is apparent that the nonlinear approach better characterizes the data in both fit, as well as being more closely aligned to the trends displayed by the observed data.

**LCS Models** To this point, the latent growth models have highlighted the different growth trends which may be estimated when individual growth across all observed time points is modeled from a linear or nonlinear perspective. The proceeding sections now introduce examples of univariate proportional change, constant change, and dual change models, as well as a bivariate dual change model.



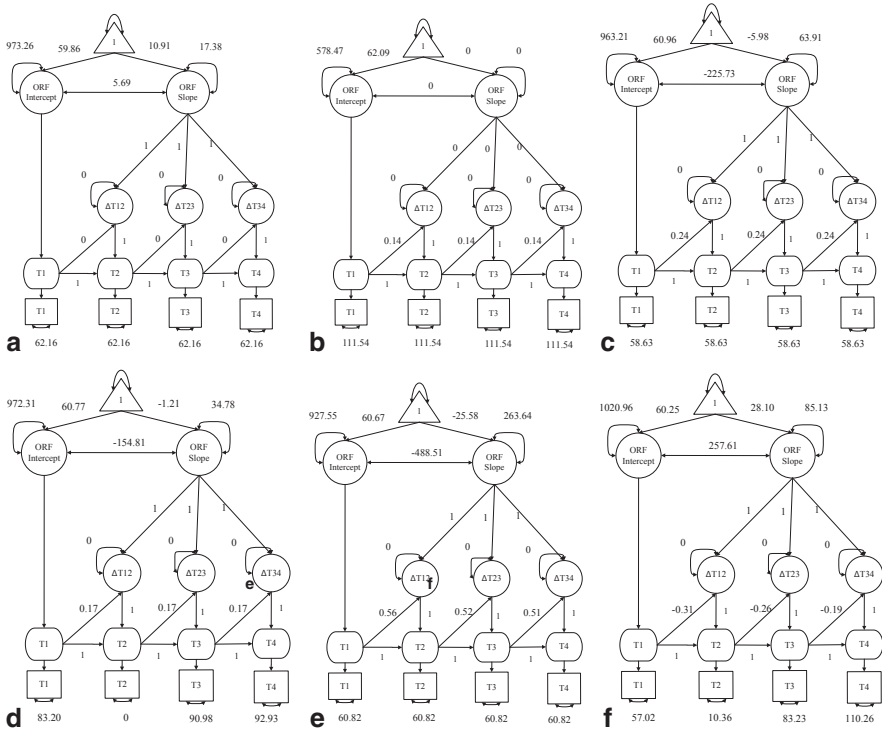
**Fig. 12.11** NWF **a** constant change, **b** proportional change, **c** dual change 1, **d** dual change 2, **e** dual change 3, and **f** dual change 4 models

Univariate LCS Models:

*NWF* The first LCS model fit to the NWF data was the constant change model where the  $\beta$  coefficients (i.e., the proportional change coefficients) were fixed to 0. Resulting fit for this model (Table 12.2) from a criteria-based evaluation was excellent [ $\chi^2(8)=7,822$ , CFI=.96, TLI=.97, RMSEA=.115 (95% CI=.113, .117)]. Note that the fit for the constant change model was identical to the fit of the linear latent growth curve model. Because the autoprotection effects were not modeled, the LCS constant change model is reduced to a linear latent growth curve model. Following this specification, the proportional change model was estimated where the  $\alpha$  coefficients (i.e., the loadings of the LCSs on the slope factor) were fixed to 0, as were the means and variances of the latent slope factor. Model fit based on the CFI and TLI met minimal thresholds for acceptable fit; however, a  $\chi^2$  difference test indicated that the constant change model provided a more parsimonious fit ( $\Delta\chi^2=2,300$ ,  $\Delta df=2$ ,  $p < .001$ ). A comparison of the model coefficients for the constant and proportional change models are provided in Fig. 12.11a, b, respectively. Note that the average NWF intercept is approximately equal across both models, as is the variance of the intercept, and the error variances across the four time points.

Specification of the first dual change score model (i.e., constrained autopportion and error variances; Fig. 12.11c), resulted in significantly better fit than the constant change score model ( $\Delta\chi^2=3,589$ ,  $\Delta df=1$ ,  $p < .001$ ), and though the BIC value for this model (2,754,233) was lower than the linear growth model (2,757,813), it was larger than the nonlinear growth model (2,752,935). Both the freed error variance-constrained autopportion (i.e., dual change 2) model and the constrained autopportion freed error variance model (i.e., dual change 3) provided better fit than the fully constrained dual change model ( $p < .001$ ; Table 12.2). When both error variances and autopportion parameters were freed for estimation (i.e., dual change 4), this model provided the most parsimonious fit to the data and fit significantly better than dual change 2 ( $\Delta\chi^2=1,734$ ,  $\Delta df=2$ ,  $p < .001$ ). Figure 12.11c, f display the resulting coefficients of each of the estimated dual score models 1–4, respectively. When the error variances were constrained but the autopportions were either constrained (Fig. 12.11c) or freed (Fig. 12.11e), it may be observed that the biggest impact of this differential specification was on not only the autopportion coefficients but also the covariance between the intercept and slope factors (i.e.,  $-343.41$  for dual change 1 and  $-542.06$  for dual change 3) and the variance of the slope factor (i.e.,  $159.93$  for dual change 1 and  $364.05$  for dual change 3). A similar phenomenon occurs when the error variances are freed and the autopportions are differentially fixed (dual change 2; Fig. 12.11d) or freed (dual change 4; Fig. 12.11f).

Differences in the model coefficients across the four specifications underscore the importance of theory relative to parameter constraints for the LCS model. Is the researcher most interested in the best fitting model? Does one assume that the error variances are the same over time, or does the nature of the assessment allow a relaxation of that constraint? Should there be an expectation that the dynamic relation is not invariant over time? As it pertains to these data, a well fitting model was desirable, yet at the same time, a strong theory did not exist as to why the error variances should be freed across the assessment periods. Previous research has shown that NWF in second grade may exhibit a slight bimodal distribution (Catts et al., 2009), thus, it is plausible that the dynamic relation would not be invariant over time. Subsequently, dual change model 3 was selected for the explication of coefficients. Similar to the nonlinear latent growth model, the average NWF score was 69 cls with an associated variance of 835.54. The mean slope for this model was negative at  $-42.09$  with a variance of 364.05. The mean of the slope factor should be interpreted with caution as it does not represent average growth in the same way that it was viewed in the linear and nonlinear growth models. Instead, the mean is interpreted as the average unique effect that contributed to the estimated LCS above the proportional change coefficient. In this model, the proportional coefficients were  $\beta = 0.67$  ( $p < .001$ ) for the effect of time 1 NWF on the change score between times 1 and 2,  $\beta = 0.74$  ( $p < .001$ ) for the effect of time 2 on the second change score, and  $\beta = 0.62$  ( $p < .001$ ) for the effect of time 3 on the third change score.



**Fig. 12.12** ORF **a** constant change, **b** proportional change, **c** dual change 1, **d** dual change 2, **e** dual change 3, and **f** dual change 4 models

Univariate LCS Models:

*ORF* The six LCS model specifications were also applied to the ORF data; yet different results were obtained compared to the NWF models. Neither the proportional change nor constant change score models fit significantly better than the nonlinear latent growth models (Table 12.2), but the constant change model retained the same fit as the linear latent growth model. A comparison of the constant and proportional change model results (Fig. 12.12a, b, respectively) highlights that the specifications yielded variances around the ORF intercept which were quite discrepant (i.e., 973.26 for constant change, 578.47 for proportional change), as well as larger residual variances for the proportional change model (111.54) compared to the constant change model (62.16). The four dual change score models (Table 12.2; Fig. 12.12c, f) showed that, of the four specifications, the freed error variance constrained autoprotection (dual change 2) and the freed error variance freed autoprotection (dual change 4) models provided the best fit when compared to the other dual change models, as well as when compared to the nonlinear model. When considering the theoretical relevance of each model, it was determined that dual change 4 would provide the best explication of the model to data relation. Catts et al. (2009) showed

that the distribution of ORF in second grade moves from a skewed distribution at the fall to a normal distribution at the spring; thus, it was plausible that the auto-proportion coefficients may change over time given the nonnormality. Moreover, the presence of nonequivalent ORF scores may have an impact on the reliability of such scores. (Ardoin & Christ, 2009; Cummings, Park, & Schaper, 2013; Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008; Petscher, Cummings, Biancarosa, & Fien, 2013). From these two potential influences on the reliability and validity of ORF scores, dual change model 4 was selected to describe the results.

The mean ORF score in the dual change ORF model was approximately 60 wcpm with an average unique slope of 28 wcpm per assessment period. The proportional change coefficient for this model and outcome was  $-0.31$  for the effect at time 1 ( $p < .01$ ),  $-0.26$  for the effect at time 2, and  $-0.19$  for the effect at time 3.

### Estimating Individual LCSs

Resulting coefficients from the NWF and ORF LCS models can subsequently be used to create individual predicted LCSs using Eq. 12.4. The LCSs for NWF would then be estimated with:

$$\begin{aligned}\Delta NWF[t]_{12} &= -42.09 + 0.67 * NWF_1 \\ \Delta NWF[t]_{23} &= -42.09 + 0.74 * NWF_2 \\ \Delta NWF[t]_{34} &= -42.09 + 0.62 * NWF_3.\end{aligned}$$

The NWF models show that from the first time point to the second, the predicted LCS simultaneously decreased by 42 points and increased proportionally by 0.67 points relative to the time 1 NWF score. The individual LCSs for ORF would be constructed with:

$$\begin{aligned}\Delta ORF[t]_{12} &= 28.10 + -0.31 * ORF_1 \\ \Delta ORF[t]_{23} &= 28.10 + -0.26 * ORF_2 \\ \Delta ORF[t]_{34} &= 28.10 + -0.19 * ORF_3.\end{aligned}$$

The ORF results show that across the assessment periods, students increased their ORF scores additively by 28.10, but decreased proportionally depending on when change was estimated. Moreover, the decreasing magnitude of the auto-proportion coefficient indicated that the impact of the autoregressive effect diminishes over time.

Although the discrepancy between the direction of the NWF and ORF auto-proportion effects may appear counterintuitive, suppose we take a student whose NWF performance at each time point was at the mean reported in Table 12.1, their estimated LCSs for each occasion would be:

$$4.28 = -42.09 + 0.67 * 69.21$$

$$12.02 = -42.09 + 0.74 * 73.12$$

$$10.35 = -42.09 + 0.62 * 84.58.$$

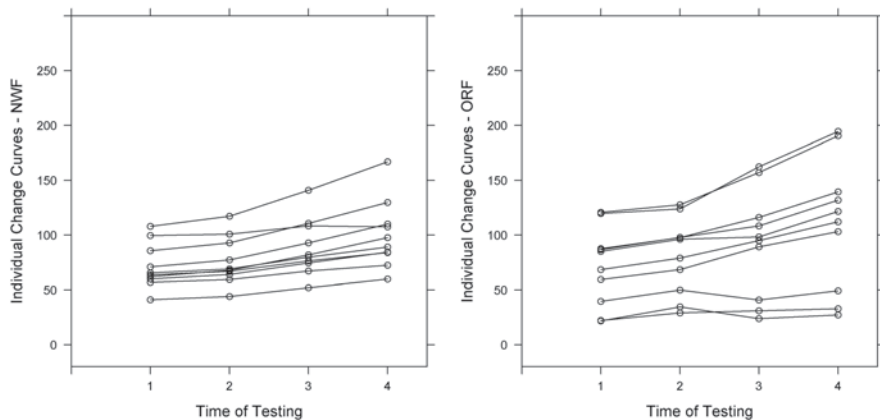
The predicted LCSs indicate that the greatest change in NWF was made between times 2 and 3 (i.e., December and February), whereas the least change occurred between times 1 and 2. Subsequently, an individual growth trajectory can be constructed from the predicted LCSs. In this example, the estimated time 1 score is 69.21, at time 2 it is 73.49 (i.e., 69.21+4.28), time 3 is 85.51 (73.49+12.02), and time 4 is 95.86 (85.51+10.35). In the same way, if we use the mean ORF score at each time point, the estimated LCSs are:

$$9.19 = 28.10 + -0.31 * 61.00$$

$$9.89 = 28.10 + -0.26 * 70.03$$

$$12.96 = 28.10 + -0.19 * 79.71,$$

which when used to calculate predicted scores for each time point, would be 61 at time 1, 70.19 at time 2 (61.00+9.19), 80.08 at time 3 (70.19+9.89), and 93.04 at time 4 (80.08+12.96). From these calculations, the expected LCSs are positive for both NWF and ORF despite the differences in the direction of the average slope and the autoprotection coefficients. This further underscores the earlier point that one must proceed with caution when interpreting the individual parameters as they represent unique contributions to the LCS. The grouped individual latent trajectories for the random sample of students on NWF and ORF are plotted in Fig. 12.13. Note that these figures appear more similar to the nonlinear growth models, yet also



**Fig. 12.13** Grouped latent change plots of non-word fluency (*NWF*) and oral reading fluency (*ORF*)

retain a distinctiveness that highlights how using the coupling effects can inform the latent trajectory different from the non-linear growth curve.

### Bivariate Dual Change Scores

The final model illustration of the LCS models is the bivariate specification which incorporates lag or coupling effects on the change scores. Prior to estimating the bivariate dual change model, a parallel process linear growth model was estimated to serve as the baseline comparison. This model estimated the growth of NWF and ORF as well as the covariances across intercepts and slopes of both measures. Resulting fit for this model fell short of acceptable criteria [ $\chi^2(28)=60,632$ , CFI=.93, TLI=.93, RMSEA=.171 (95% CI=.170, .172), BIC=5,120,796]. Conversely, the bivariate LCS model provided good fit to the data based on criterion indices [ $\chi^2(17)=18,438$ , CFI=.98, TLI=.96, RMSEA=.121 (95% CI=.120, .123)], and the BIC was lower (5,078,692) compared to the parallel process model (5,120,796) suggesting the former model is more parsimonious. The resulting parameters for the bivariate LCS model are provided in Fig. 12.14.

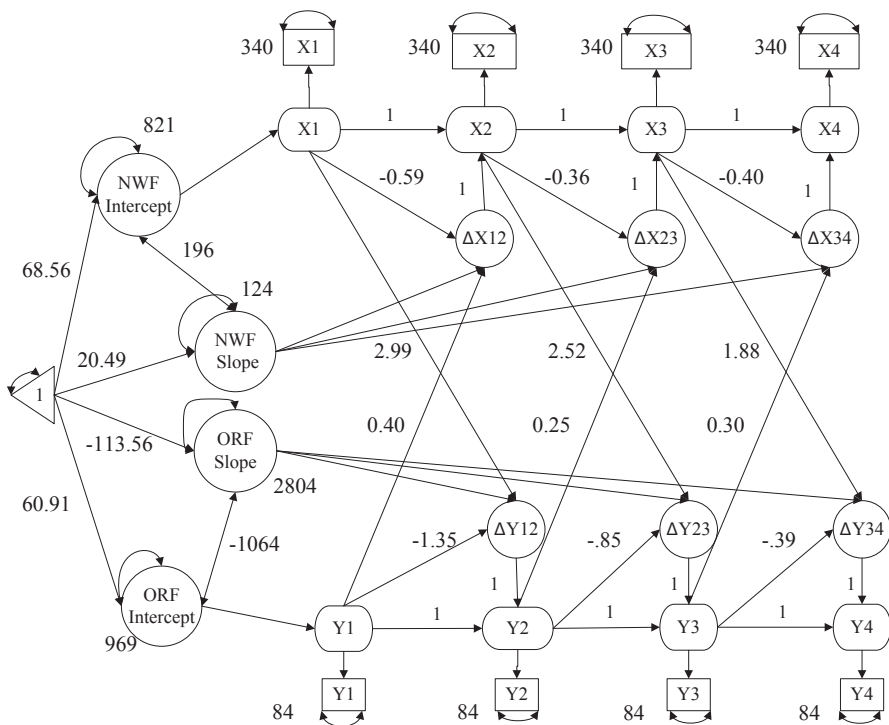


Fig. 12.14 Bivariate dual change score model parameters



Similar to the univariate dual change models, the estimated parameters included means for the latent intercepts (68.56 and 60.91 for NWF and ORF) and slopes (20.49 and -113.56 for NWF and ORF) as well as the variances, covariances, and proportional change coefficients. The inclusion of coupling effects of NWF on ORF change as well as ORF on NWF change demonstrated the differential contributions each makes to the estimated LCSs. Model coefficients from Fig. 12.13 can be inserted into Eq. 12.5 to create the estimated scores for NWF and ORF as:

$$\Delta NWF[t]_{12} = 20.49 - 0.59 * NWF_1 + 0.40 * ORF_1$$

$$\Delta NWF[t]_{23} = 20.49 - 0.36 * NWF_2 + 0.25 * ORF_2$$

$$\Delta NWF[t]_{34} = 20.49 - 0.40 * NWF_3 + 0.30 * ORF_3$$

and

$$\Delta ORF[t]_{12} = -113.56 - 1.35 * ORF_1 + 2.99 * NWF_1$$

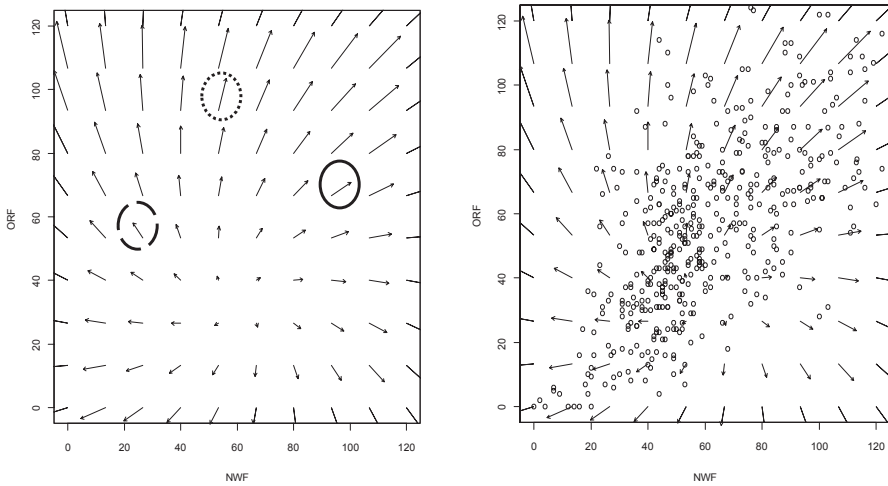
$$\Delta ORF[t]_{23} = -113.56 - 0.85 * ORF_2 + 2.52 * NWF_2$$

$$\Delta ORF[t]_{34} = -113.56 - 0.39 * ORF_3 + 1.88 * NWF_3.$$

From these equations several important implications can be seen. When considering NWF, both the proportional change and coupling coefficients were statistically significant, indicating that NWF was a leading indicator of NWF change, and also that ORF was a leading indicator (i.e., determinant) of change in NWF. The autopropportion coefficients for NWF varied across the change scores, with a much larger, negative effect occurring as a function of time 1 NWF on the first change score (-.59) compared to the effect of time 2 NWF on the second change score (-.36). Because the direction of the coefficients is negative, the interpretation of the model is such that while students make an average gain of 20.49 between assessments, there is simultaneously a negative proportional effect such that students who have the lowest fluency scores at the previous time point changed the most. Additionally, given the average change and proportional change effects, the positive coefficient for ORF indicated that the unique coupling effect was the strongest for individuals with higher ORF scores (i.e., those with higher ORF scores changed the most on NWF).

ORF as an outcome demonstrated a similar pattern of coefficients for the autopropportion and coupling effects. Across the change score equations, the amount of average change was negative, as were the autopropportion coefficients. This suggested that individuals with lower ORF scores in prior assessments changed the most, after controlling for the average effect. Along with this coefficient, students with higher, previous NWF scores also experienced the greatest change in ORF at any time point.

As with the previously defined models, graphs are the most helpful utility to facilitate understanding the effects of the LCS models, yet the bivariate nature of



**Fig. 12.15** Vector plots of non-word fluency (*NWF*) and oral reading fluency (*ORF*)

the model requires a slightly different mechanism to present results. In this way, vector field plots are advantageous in evaluating joint latent trajectories. The arrows in Fig. 12.15a represent the initial values of each NWF–ORF combination, and the direction of the arrow shows the type of expected change to occur from the initial status combinations in the vector field. For example, the arrow enclosed by the solid circle depicts the expected change in NWF and ORF where an individual started the year with an NWF score of approximately 100 and an ORF score of 70. Based on the direction of the arrow, the individual is predicted to grow positively in both skills, and the expectation is that greater change will occur for NWF compared to ORF. Conversely, the dotted circle identifies an individual who starts the year with a NWF score of approximately 60 and an ORF score of 90. With the arrow pointing up, the vector suggests that the predicted magnitude of change is much stronger for ORF, but remains rather stagnant for NWF. Finally, the dashed circle represents an individual who scored low on NWF (approximately 25 c/s) relative to ORF (approximately 60 wcpm) and the direction of the vector indicates that positive change is expected for ORF while negative change is expected for NWF.

A cursory evaluation of the plot might lead one to be surprised that so many individuals have negative expected change scores, but it is important to note that the vector field plot provides information on expected change scores based on the model equations, yet it may not reflect how students in the sample actually performed. Figure 12.15b overlays a scatterplot for a random selection of 500 individuals in the sample. The density of the scatter can better facilitate which of the arrows in the vector field reflect actual observed performance. When comparing Figs. 12.15a, b, the implication of the scatterplot is that most students are, in fact, predicted to have positive LCSs for both measures, but there are certainly a number of individuals who are predicted to either grow positively on just one of the assessments or negatively change on both.

## Summary

The goal of this chapter was to introduce the reader to LCS modeling and highlight its complexities and the information it yields in the context of traditional latent growth models. Modeling individual trajectories in educational research presents several complex issues that have previously been evaluated using either structural or dynamic modeling. Each has its set of distinct advantages and disadvantages, yet the strengths of both may be combined in the LCS model. Such a framework allows one to be able to understand the nature of change for a given assessment as well as the determinants of such change (McArdle & Grimm, 2010).

A distinct advantage to using the LCS model with fluency research is that it allows one to disaggregate growth into multiple change scores which can be useful in isolating where a student is most likely to change the most. The model coefficients from Fig. 12.13 highlight that the causal portions of the model are useful in understanding the effects on change. While the estimates for NWF varied slightly, a much clearer separation of magnitude was observed for ORF where the estimated autopropotion effect of time 1 was  $-1.35$  compared with the  $-.39$  effect on time 3. The finding of discrepant effects could not have been ascertained in a traditional latent growth model.

Moreover, not only are the dynamic growth and causal portions of the model useful to understanding change but also the coupling effects may assist in identifying which predictors yield the most useful information about change. In the current illustration it was found that prior NWF performance was the strongest predictor of change in ORF, despite the strong effect of the ORF autopropotion coefficient. This finding may assist in yielding new understanding about reciprocal causation when predicting individual differences in fluency type outcomes.

## Extensions

The univariate and bivariate examples both illustrate how the LCS model can be applied to data when there is one measured variable per construct at each measurement occasion. Researchers frequently have access to multiple measures of a given construct such as multiple passages for ORF, which allows a common factor to be estimated as a function of the shared variance across the individual passages. McArdle and Prindle (2008) used a common-factor LCS model to evaluate the impact of cognition training on the elderly. Similarly, Calhoun and Petscher (2013) used a common-factor LCS model to test the impact of different reading interventions for middle and high school students. A notable difference between the examples presented here and the common-factor LCS model is that there is a strict requirement for invariance of the factor loadings across the measurement occasions (McArdle & Hamagami, 2001). As with many longitudinal applications of structural equation modeling (SEM) using common factors, strict invariance is often not tenable (Millsap, 2012), with many models meeting requirements for partial measurement invariance, thus it is important that one carefully evaluates the invariance of the loadings prior to proceeding with a common-factor LCS model.

Additional aspects of the LCS model that may be useful for researchers working with fluency data are multiple-group, multilevel, and multivariate change models. As with many SEM models, questions about invariance of means, variances, and loadings are of interest when one has collected data where multiple groups are involved. Several studies have used the multiple group approach to evaluate how different autopropotion and coupling effects differ between males and females (McArdle & Grimm, 2010) as well as whether an intervention had a different effect in the treatment or control groups (Calhoun & Petscher, 2013; McArdle & Prindle, 2008). A final extension worth noting is that the LCS model can be fit in a multilevel framework. The presented illustration focused on fitting the model when students were the unit of interest, yet it is possible to extend this model to examine the coefficients when considering nested structures such as students in classrooms or schools (e.g., Petscher, 2012).

## Final Thoughts

This chapter has illustrated the LCS as a potential alternative to traditional linear and nonlinear growth modeling as it simultaneously models individual growth as well as determinants of change. Because fluency data may often retain distributional properties which may restrict individual differences, or more appropriately, may mask differences from estimated means effects, growth models may be inefficient as capturing the developmental nature of change. As shown in this example, the LCS model demonstrated much greater model parsimony to the data compared to the growth models, and displayed predicted individual growth curves from the change scores which were not too dissimilar from the observed fluency scores or the predicted nonlinear individual growth curves. The flexibility to fit dual change, constant change, or proportional change models allows for researchers with fluency panel data to potentially obtain a richer understanding of change over time and it is our hope that these models will allow users to better study individual differences in fluency development.

## References

- Al Otaiba, S., Petscher, Y., Pappamihiel, N. E., Williams, R. S., Drylund, A. K., & Connor, C. M. (2009). Modeling oral reading fluency development in Latino students: A longitudinal study across second and third grade. *Journal of Educational Psychology, 101*, 315–329.
- Arbuckle, J. L. (2006). AMOS (Version 7.0). Chicago: SPSS.
- Ardoin, S. P., & Christ, T. J. (2009). Curriculum based measurement of oral reading: Estimates of standard error when monitoring progress using alternate passage sets. *School Psychology Review, 38*, 266–283.
- Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (ALT) models: A synthesis of two traditions. *Sociological Methods and Research, 32*, 336–383.
- Bowles, R. P., & Montroy, J. J. (2013). Latent growth curve modeling using structural equation modeling. In Y. Petscher, C. Schatschneider, & D. L. Compton (Eds.), *Applied quantitative analysis in education and the social sciences* (pp. 265–303). New York: Routledge.

- Branum-Martin, L. (2013). Multilevel modeling: Practical examples to illustrate a special case of SEM. In Y. Petscher, C. Schatschneider, & D. L. Compton (Eds.), *Applied quantitative analysis in the social sciences* (pp. 95–124). New York: Routledge.
- Calhoun, M. B., & Petscher, Y. (2013). Individual sensitivity to instruction: Examining reading gains across three middle school reading projects. *Reading and Writing, 26*, 565–592.
- Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on early identification. *Journal of Learning Disabilities, 42*, 163–176.
- Crowe, E. C., Connor, C. M., & Petscher, Y. (2009). Examining the core relations between poverty, reading curriculums, and first through third grade reading achievement. *Journal of School Psychology, 47*, 187–214.
- Cummings, K. D., Dewey, B., Latimer, R., & Good, R. H. (2011). Pathways to word reading and decoding: The roles of automaticity and accuracy. *School Psychology Review, 40*, 284–295.
- Cummings, K. D., Park, Y., & Schaper, H. A. B. (2013). Form effects on DIBELS next oral reading fluency progress-monitoring passages. *Assessment for Effective Intervention, 38*, 91–104.
- Fletcher, J. M., Coulter, W. A., Reschly, D. J., & Vaughn, S. (2004). Alternative approaches to the definition and identification of learning disabilities: Some questions and answers. *Annals of Dyslexia, 54*, 304–331.
- Foorman, B. R., Petscher, Y., Lefsky, E. B., & Toste, J. R. (2010). Reading first in florida: Five years of improvement. *Journal of Literacy Research, 42*(1), 71–93. doi:http://dx.doi.org/10.1080/10862960903583202.
- Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' ORF using DIBELS. *Journal of School Psychology, 46*, 315–342.
- Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development, 58*, 80–92.
- Good, R. H., & Kaminski, R. A. (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene: Institute for the Development of Educational Achievement.
- Good, R. H., Kaminski, R. A., Smith, S., Laimon, D., & Dill, S. (2001). *Dynamic indicators of basic early literacy skills* (5th ed.). Eugene: University of Oregon.
- Good, R. H., Wallin, J., Simmons, D. C., Kameenui, E. J., & Kaminski, R. A. (2002). *System-wide percentile ranks for DIBELS benchmark assessment (Technical Report, No. 9)*. Eugene, OR: University of Oregon
- Grimm, K. J. (2012). Intercept centering and time coding in latent difference score models. *Structural Equation Modeling: A Multidisciplinary Journal, 19*, 137–151.
- Grimm, K. J., & Widaman, K. F. (2010). Residual structures in latent growth curve modeling. *Structural Equation Modeling, 17*, 424–442. (Mplus Scripts).
- Hertzog, C., & Nesselroade, J. R. (1987). Beyond autoregressive models: Some implications of the trait-state distinction for the structural modeling of developmental change. *Child Development, 58*, 93–109.
- Hox, J. J. (2000). *Multilevel analyses of grouped and longitudinal data*. Mahwah: Lawrence Erlbaum Associates Publishers.
- Joreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*, 239–251.
- Joreskog, K. G., & Sorbom, D. (1979). *Advances in factor analysis and structural equation models*. New York: University Press of America.
- Joreskog, K. G., & Van Thillo, M. (1972). *LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables*. Princeton: Educational Testing Service.
- Kim, Y. S., Petscher, Y., Schatschneider, C., Foorman, B. R. (2010). Does growth in oral reading fluency matter in reading comprehension achievement? *Journal of Educational Psychology, 102*, 652–667.
- Malone, P. S., Lansford, J. E., Castellino, D. R., Berlin, L. J., Dodge, K. A., Bates, J. E., & Peettit, G. S. (2004). Divorce and child behavior problems: Applying latent change score models to life event data. *Structural Equation Modeling, 11*, 401–423.

- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577–605.
- McArdle, J. J., & Grimm, K. J. (2010). Five steps in latent curve and latent change modeling with longitudinal data. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Longitudinal research with latent variables* (pp. 245–274). New York: Springer.
- McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analysis with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change*. Washington, D.C.: American Psychological Association.
- McArdle, J. J., & Nesselroade, J. R. (1994). Structuring data to study development and change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological innovations* (pp. 223–267). Mahwah: Erlbaum.
- McArdle, J. J., & Prindle, J. J. (2008). A latent change score analysis of a randomized clinical trial in reasoning training. *Psychology and Aging*, *23*(4), 702–719. doi:http://dx.doi.org/10.1037/a0014349.
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equation models. *Psychological Methods*, *10*, 259–284.
- Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: Latent growth curves with ordinal outcomes. *Psychological Methods*, *9*, 301–333.
- Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. New York: Taylor and Francis.
- Muthén, B. (2004). Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 345–368). Thousand Oaks: Sage Publications, Inc.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus User's Guide*. (7th ed.). Los Angeles: Muthén & Muthén.
- O'Connell, A. A., & McCoach, D. B. (2008). *Multilevel modeling of educational data*. Charlotte: Information Age Publishing, Inc.
- O'Connell, A. A., Logan, J. A. R., Pentimonti, J. M., & McCoach, D. B. (2013). Linear and quadratic growth models for continuous and dichotomous outcomes. In Y. Petscher, C. Schatschneider, & D. L. Compton (Eds.), *Applied quantitative analysis in education and the social sciences* (pp. 125–168). New York: Routledge.
- Petscher, Y. (2012). *Estimating multivariate and multilevel latent change scores*. Paper presented at the Society for Research on Child Development, Tampa, Florida.
- Petscher, Y., & Kim, Y. S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology*, *49*, 107–129.
- Petscher, Y., Cummings, K. D., Biancarosa, G., & Fien, H. (2013). Advanced (measurement) applications of curriculum-based measurement in reading. *Assessment for Effective Intervention*, *38*, 71–75.
- Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology*, *25*, 111–196.
- Reynolds, M. R., & Turek, J. (2012). A dynamic developmental link between verbal comprehension knowledge (Gc) and reading comprehension: Verbal comprehension-knowledge drives positive change in reading comprehension. *Journal of School Psychology*, *50*, 841–863.
- Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Not just speed reading: Accuracy of the DIBELS oral reading fluency measure for predicting high-stakes third grade reading comprehension outcomes. *Journal of School Psychology*, *46*, 343–366.
- Sivo, S. A., Fan, X., & Witte, E. L. (2005). The biasing effects of unmodeled ARMA time series processes on latent growth curve model estimates. *Structural Equation Modeling*, *12*, 215–232.
- Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology*, *93*, 735–749.
- Stoel, R. D., Van den Wittenboer, G., & Hox, J. J. (2004). Methodological issues in the application of the latent growth curve model. In K. van Montfort, H. Oud, & A. Satorra (Eds.), *Recent developments on structural equation modeling: Theory and applications* (pp. 241–262). Amsterdam: Kluwer Academic Press.
- Zhiyong, Z., McArdle, J. J., Hamagami, A., & Grimm, K. (2013). *RAMpath: Structural equation modeling using RAM notation* (R package version 0.3.6).

# Chapter 13

## Conclusion: Oral Reading Fluency or Reading Aloud from Text: An Analysis Through a Unified View of Construct Validity

Christine A. Espin and Stanley L. Deno

The chapters in this book focus on the role of fluency in the measurement of performance and progress within different academic areas. In this chapter, we reflect upon the extent to which the construct *fluency* plays a role in the validity of the scores generated by measures in academic areas. We focus specifically on the use of fluency measures within a Curriculum-Based Measurement (CBM) approach, and describe the ways in which different validity arguments reflect different proposed interpretations and uses. Key to the discussion is whether fluency is the construct being measured or whether it is a construct being used to create measures that produce technically adequate scores. To illustrate, we begin the chapter with a multiple-choice question.

The oral reading fluency measure (ORF):

- A. produces scores that measure reading fluency
- B. produces scores that measure general reading proficiency
- C. produces scores that are indicators of general reading proficiency
- D. should be referred to as a reading aloud measure (RA)
- E. all of the above
- F. none of the above

To answer this multiple choice question, one must know what the intended interpretations and uses of the scores generated by ORF are. To answer this question, one must get at the heart of validity.

---

C. A. Espin (✉)  
Leiden University, Leiden, South Holland, The Netherlands  
e-mail: [espinca@fsw.leidenuniv.nl](mailto:espinca@fsw.leidenuniv.nl)

S. L. Deno  
University of Minnesota, Minneapolis, MN, USA  
e-mail: [denox001@umn.edu](mailto:denox001@umn.edu)

© Springer Science+Business Media, LLC 2016  
K. D. Cummings, Y. Petscher (eds.), *The Fluency Construct*,  
DOI 10.1007/978-1-4939-2803-3\_13



## What is Validity?

The concept of *validity* has slowly evolved and changed over time (Hubley & Zumbo, 2011; Kane, 2013; Messick, 1989b; Moss, 2007). For example, some of us may remember learning that validity was “the extent to which a test measures what it is supposed to measure,” and that there were three types of validity—*content-*, *criterion-*, and *construct-related* validity. However, these simple definitions and descriptions no longer capture current conceptions of validity.

Current thinking about validity includes the following tenets:

1. *Validity is not a property of a test or measurement instrument.* Validity does not apply to the measurement instrument itself, but to the scores produced by measurement instrument. Specifically, it applies to the interpretations and proposed uses of the scores produced by a measurement instrument.
2. *Validity is a unified concept.* There are not different types of validity, but different types of evidence that support validity. All of validity is construct validity.
3. *Validity is a matter of degree.* Validity is not an all or nothing affair, but a matter of degree. Validation is an ongoing process of scientific inquiry. Validity may change over time as evidence accumulates or uses change.
4. *Validation should be guided by interpretation/use arguments* (IUAs). The process of validation should be guided by a coherent and complete argument outlining proposed interpretations and uses of the scores produced by the measurement instrument.

### ***Validity Is Not a Property of a Test or Measurement Instrument***

Validity is not a property of a test but a property of the scores produced by a test or measurement instrument, and it includes both the interpretations and proposed uses of those scores. Messick (1989b) defines validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment” (italics not added; p. 13). The *Standards for Educational and Psychological Testing*, a document jointly developed by The American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) to guide the development and use of tests, defines validity as the “degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (AERA et al., 1999, p. 9). The validity of a score, then, depends on the intended use of that score for decision-making, and different intended uses require different validity arguments to be made.

## ***Validity Is a Unified Concept***

**All of Validity Is Construct Validity** There are not different types of validity, but different types of evidence that support validity—and all of validity is construct validity. The view of validity as a unified concept is perhaps one of the most significant developments in recent conceptualizations of validity. As early as 1957, Loevinger described validity as a unified concept, arguing that “since predictive, concurrent, and content validities are all essentially *ad hoc*, construct validity is the whole of validity from a scientific point of view” (p. 636); however this view was not widely held at the time. Between 1954 and 1974, for example, the *Standards for Educational and Psychological Testing* described three different types of validity: content, criterion, and construct (Cizek, 2012; Moss, 2007). It was not until the 1974 and 1985 versions that the *Standards* began to incorporate the view of validity as a unified concept (Kane, 2013; Moss, 2007).

The unified view of validity coalesced with the writings of Messick (1989a, 1989b, 1995), who emphasized that “construct validity may ultimately be taken as the whole of validity in the final analysis” (Messick, 1989a, p. 21). Messick (1989a, 1989b) described validity in terms of both score interpretations and score uses. Score interpretations referred to the inferences that were to be made from the scores, and different types of inferences required different types of evidence. The types of inferences to be made depended on the potential score use. Messick (1989a, 1989b, 1995) considered both the evidential and consequential basis of validity, arguing that one had to be concerned not only with the degree of evidence supporting particular score interpretations and uses, but also with the potential consequences associated with those interpretations and uses.

Messick was not the first to refer to the process of validation as a scientific inquiry, nor the first to propose that validity was a unified concept. As mentioned earlier, Loevinger (1957), as well as others (e.g., Anastasi, 1986; Cronbach, 1971; Cronbach & Meehl, 1955; Loevinger, 1957) contributed to the development of these ideas. However, a unique aspect of Messick’s conceptualization of validity was the inclusion of the *consequences* of score interpretations (Kane, 2013). Although there is some disagreement as to whether the term *validity* should be so broadly defined (e.g., see Borsboom & Mellenbergh, 2004; Hood, 2009; Lissitz & Samuelsen, 2007), there is widespread agreement that both the evidential and consequential basis of validity should at least be taken into account during the development and use of assessments (Cizek, 2012). In the following sections, we describe in more detail what is meant by evidential and consequential basis of validity.

**Evidential Basis of Validity** The evidential basis of validity refers to the various types of evidence that might be called in to play to support validity. There are various sources of validity evidence possible; the sources selected depend on the potential score interpretations and uses (Messick, 1989a, 1989b).

The types of evidence are grouped somewhat differently from document to document, but in general they include content, internal structure, response processes, and relations to other variables (Messick, 1989b; Moss, 2007; AERA et al., 1999).

## Content Evidence

Content evidence reflects the extent to which the test content is representative of the content domain (Messick, 1989a, 1989b; Reynolds & Livingstone, 2012). Content evidence is established primarily via systematic analysis and expert opinion about the content domain and the test items that represent that domain.

## Internal Structure

Internal structure refers to the extent to which test items align with the construct the test is designed to measure (AERA et al., 1991; Reynolds & Livingstone, 2012). For example, if a construct is hypothesized to include multiple dimensions, the test can be examined to determine whether the internal structure of the test reflects those dimensions. Examination of the internal structure of a test often involves the use of factor analyses techniques.

## Response processes

Response processes are the processes underlying item or task performance (Messick, 1989b). Evidence related to response processes reflects the fit between actual responses of the examinees and the construct being measured (AERA et al., 1999; Reynolds & Livingstone, 2012), and ties scores on the measure to the theoretical rationales underlying the construct. Examination of response processes might involve techniques such as protocol analysis, computer modeling, response time analysis, and measurements of eye movement (Messick, 1989b).

## Relations to other variables

Relations to other variables refers, to the patterns of relations between scores on the measurement instrument and other variables (the criteria) thought to represent the construct being measured (AERA et al., 1999; Reynolds & Livingstone, 2012). Evidence of the relations to other variables reveals the extent to which scores from the measure fit with theoretical conceptualizations about the measure (Messick, 1989b; AERA et al., 1999). Examples of methods for establishing evidence based on relations to other variables include *test-criterion evidence*, *convergent and discriminant evidence*, *group differences/changes over time*, and *responsiveness of scores to experimental treatment* (Messick, 1989b; Reynolds & Livingstone, 2012; AERA et al., 1999).

Test-criterion evidence reflects the relation between scores on an instrument and scores on other measures representative of the construct. Test-criterion evidence is developed through both concurrent and predictive studies. Concurrent studies examine the relation between scores when the test and criterion are administered at the same time; predictive studies examine the relation between

scores when the criterion is administered later than the test (Reynolds & Livingstone, 2012).

Convergent and discriminant evidence reflect the extent to which correlations between scores on the test and other measures of the same construct are larger than correlations between scores on the test and measures of a different construct. An elegant way to examine convergent and discriminant evidence simultaneously is the formulation of a multitrait-multimethod matrix, in which evidence about multiple traits measured with multiple methods is examined (Campbell & Fiske, 1959; Messick, 1989b, Reynolds & Livingston, 2012).

Studies of group differences examine whether scores for groups differ in expected directions based on theoretical understanding of the construct. For example, we would expect scores on a test of reading competence to be higher for students in grade 6 than for students in grade 2 (Messick, 1989b; Reynolds & Livingston, 2012). In addition, longitudinal studies of score changes over time evaluate whether the scores change in expected ways based on a theoretical understanding of the construct (Messick, 1989b). For example, we would expect scores on a test of reading competence to increase for students as they move from grades 2 to 6.

Responsiveness of scores to experimental treatment refers to whether scores change in theoretically predictable ways in response to experimental manipulations. For example, we would expect that scores on a test of reading competence would increase for students following implementation of an effective reading intervention.

**Consequential Basis of Validity** The consequential basis of validity refers to consideration of the consequences of test interpretation and use (Cizek, 2012; Hubley & Zumbo, 2011; Messick, 1989a, 1989b; Reynolds & Livingston, 2012). Messick (1989b) emphasized that the consequential basis of validity does not refer to the consequences associated with *misuse* of a measurement instrument, but rather to the consequences associated with the *appropriate use* of the instrument. Under consequential validity, both score interpretation and score use are considered (Hubley & Zumbo, 2011; Messick 1989a, 1989b).

Under *score interpretation*, Messick (1989b) includes the values associated with the labels, the theories, and ideologies associated with a construct. Labels carry meaning, and the selection of the label bears on the consequences associated with use of that label. Consider, for example, potential differences in the consequences associated with the following labels used to describe severe reading difficulties: dyslexia, learning disabilities (LD), reading difficulties. Each label has a different set of values or connotations associated with it.

Related to the values associated with labels are the values associated with the theories undergirding the construct being measured. Messick (1989b) argues that scores are interpreted within the theoretical framework of the person viewing the scores. For example, a score on a reading competence test will carry different meaning when interpreted from the point of view that reading difficulties are caused by neurological impairments rather than from the point of view that reading difficulties are caused by poor instruction. Such differences in interpretations might lead to different consequences for the examinee.

Finally, scores are interpreted within the ideologies held by the examiner. For example, consequences of score interpretations will differ for someone who believes

that children with disabilities have a “right” to be educated in the same setting as children without disabilities than for someone who believes that children with disabilities have a “right” to be educated in a special school with trained specialists.

With regard to *score use*, Messick (1989b) argues that the potential social consequences of score use must be considered, including both intended and unintended consequences: “The central question is whether the proposed testing *should* serve as a means to the intended end, in light of other ends it might inadvertently serve ...” (italics not added, p. 85). Under social consequences of testing, Messick includes consideration of the use of the test compared to not testing at all or to using an alternative approach, and the side effects and by-products associated with testing.

Suppose, for example, that a test of verbal ability consistently produces lower scores for males than for females. The first step is to ensure that all potential sources of test invalidity have been ruled out, including construct underrepresentation and construct irrelevant test variance (Hubley & Zumbo, 2011; Messick, 1989b; Reynolds & Livingston, 2012). *Construct underrepresentation* occurs when a construct is not fully represented on the test. For example, a test in mathematics that includes only multiplication problems would underrepresent the construct “mathematics ability.” *Construct irrelevant variance* occurs when test scores are influenced by characteristics, content, or skills that are unrelated to the construct being measured. For example, a test of mathematics ability that requires a large amount of reading would produce scores that reflect not only mathematics but also reading abilities.

If potential sources of invalidity have been ruled out as an explanation for differences in group scores, then one might assume that the scores reflect true group differences. The second step is then to consider the social consequences associated with the use of such test scores. Judgments about social consequences rely on the values held by society, the definition of the construct, and the potential uses of the scores for decision-making.

To return to our example, let us assume that scores on the test of verbal ability are used to make decisions about placement of students into special education. What are the implications for the use of such test scores for making placement decisions, especially if we see that males get referred more often to special education than females? Is such an outcome desirable? One could argue that it is a desired outcome because male students are getting help, and they are in greater need of such help than female students. On the other hand, one could argue that male students are unnecessarily and unfairly being placed in special education, a decision that could potentially affect the rest of their lives. Which answer is correct? It is probably obvious that both responses might be considered to be correct. The social consequences associated with the use of test scores can be seen as both positive and negative (Hubley & Zumbo, 2011).

The example of the test of verbal ability illustrates the reciprocal relation between the evidential and consequential basis of validity. The social consequences associated with the use of test scores both rely on and influence the evidential basis of validity. It is this reciprocal relation that led Messick (1989b) to argue for a unified concept of validity: “The value implications of score interpretation are not just a part of score meaning, but a socially relevant part that often triggers score-based actions and serves to link the construct measured to questions of social policy” (Messick, 1989b, p. 63).

### ***Validity Is a Matter of Degree***

Validity is not an all or nothing affair. It is a matter of degree. Validation is an ongoing process of scientific inquiry. Thus, it is not appropriate to claim that a test score is valid or not valid, but only to describe the extent to which the existing data support the interpretations of the test score for particular uses (Messick, 1989a, 1989b; AERA et al., 1999). It is important to make an integrated, evaluative judgment of the degree to which the evidence supports score interpretation and use, and the potential consequences of such score interpretation and use (Messick, 1989a, 1989b). Each score interpretation and use must be validated (Reynolds & Livingston, 2012; Kane, 2013). Judgment about the validity of a score is made in part by comparing use of that score to the use of scores produced by other instruments, or to use of no at all (Messick, 1989b).

Accumulating evidence may support or call into question the validity of particular score interpretations and uses. If support is not found, then either the assessment instrument or the theory underlying the assessment instrument must change.

### ***Validation Should be Guided: Interpretation/Use Arguments (IUs)***

The process of validation should be guided by a coherent and complete argument that lays out the proposed interpretations and uses of the scores (Cronbach, 1988; Kane, 1992, 2006, 2013). An argument-based approach to validation provides a framework for organizing and evaluating the claims made about test scores and uses (Cronbach, 1988; Kane, 2013). Kane (2013) states that the core idea of an argument-based approach to validation is “to state the proposed interpretation and use explicitly and in some detail, and then to evaluate the plausibility of these proposals” (p. 1). An argument-based approach to validation includes two steps: first state the claims, and second evaluate those claims (Kane, 2013). Validation, then, can be seen as an “empirical evaluation of the meaning and consequences of measurement” (Messick, 1995, p. 747). As interpretation/uses of scores change over time, validity changes over time. New arguments must be built, and new evaluations of the evidence must be made (Kane, 2013). The more ambitious the proposed interpretations/and uses the greater the empirical evidence needed to support those interpretations and uses (Kane, 2013).

### **Validity of Scores from Measures Involving Fluent Reading**

In the first part of the chapter, we outlined four tenets that capture current thinking about the concept of validity. In the second section of the chapter, we focus on scores from two specific fluency-based measures, both used to monitor student progress in reading within a curriculum-based measurement (CBM) system: reading aloud and maze selection. For reading aloud, students read aloud from a text for

1 min and the number of correctly read words is tallied. We refer to this score as words read correctly (WRC). In maze selection, students read silently from a text in which every 7th word is deleted and replaced with a multiple-choice item. Students read for 1 to 3 min selecting words as they read, and the number of correct choices is tallied. We refer to this score as correct maze choices (CMC).

In a CBM system, students are measured frequently (e.g., weekly) over time using reading aloud or maze passages. WRC or CMC scores are placed on a graph to represent growth and progress in general reading proficiency. Data are used to evaluate the effects of instruction on student progress. If the data reveal a lack of progress, it is a signal that instruction needs to be changed (Deno, 1985; Deno & Fuchs, 1987).

Over the years, much of the research on CBM in reading has focused on examining the reliability and validity of scores produced by the reading aloud and maze selection measures (Wayman, Wallace, Wiley, Tichá, & Espin, 2007). In the sections that follow we illustrate how each tenet of validity applies to these two scores. Our intent is not to provide a comprehensive review of the CBM literature on reading (for reviews, see Espin & Tindal, 1998; Marston, 1989; Reschly, Busch, Betts, Deno, & Long, 2009; Stecker, Fuchs, & Fuchs, 2005; Wayman et al., 2007; Yeo, 2010, 2011), but merely to reflect upon the research within the framework of current thinking about validity.

### ***Validation of WRC and CMC Scores Should be Guided by IUAs***

The building of an interpretation and use argument (IUA) comes both at the beginning and the end of the validation process. By stating the IUA at the beginning of the validation process, one has a framework against which to judge the extent to which the data support the validity of the scores for a particular interpretation and use. While the original IUA for developing CBM reading scores was laid out by Deno (1985), more recent IUAs for CBM are more often implied than specified. It would be beneficial for CBM researchers and test developers to lay out the specific arguments for various interpretations and uses of CBM scores in order to allow for a systematic evaluation of the degree of empirical support for such interpretations and uses. We return to the IUA again at the end of this section.

### ***Validity Is Not a Property of Reading Aloud or Maze Selection Measures but of the Scores (WRC and CMC) Produced by the Measures***

Validity is not a property of the measure, but of the scores produced by the measure. It is not the validity of the reading aloud or maze selection measures, but the validity of WRC and CMC—the scores produced by the measures—that is of interest. Moreover, it is not the scores themselves, but the interpretations and uses of the scores that are of concern in a validity argument.



With regard to score interpretation, within a CBM context WRC is not interpreted as an indicator of the construct *reading fluency*, but rather the construct *general reading proficiency* (Deno & Marston, 2006). Likewise, CMC is not interpreted as an indicator of the construct *reading comprehension*, but rather the construct *general reading proficiency*. If one follows this logic, then the use of the term ORF is somewhat of a misnomer for the CBM reading aloud score because it implies that score interpretation relates to the construct reading fluency. Granted, the number of WRC in 1 min could be used as an indicator of the construct reading fluency, but then the interpretation and use argument built around the scores would need to be different (e.g., see Valencia, Smith, Reece, Li, Wixson, & Newman, 2010). Within a CBM context, however, WRC and CMC are used as indicators of general reading proficiency; thus, interpretation/use arguments are built around the extent to which the measures relate to broad measures of reading proficiency. Such measures include both fluency and comprehension.

With regard to score use, the way in which CBM reading scores have been, and are being, used has changed considerably over time. CBM was originally designed to be used in a formative assessment approach by special teachers to systematically evaluate the effects of their instruction with students who had reading disabilities (Deno, 1985; Deno & Fuchs, 1987). Most recently it has been used as a part of a response-to-intervention (RTI) approach to identify students with learning disabilities (LD) and to make placement decisions for those students. (e.g., see Fuchs, Mock, Morgan, & Young, 2003; Jimerson, Burns, & VanDerHeyden, 2007; Marston, Muyskens, Lau, & Canter, 2003; Speece, Case, & Molloy, 2003). Although the interpretation of CBM scores remains somewhat similar across these two uses (the scores are meant to reflect general reading proficiency for both uses), the use of the scores, and the values and consequences associated with the interpretations and uses, vary dramatically. Each set of uses and interpretations requires a different IUA, and each IUA requires different sources and standards of evidence (for examples of types of evidence specific to RTI see Fuchs & Deshler, 2007; Fuchs, 2003; Jimerson et al., 2007; Vaughn, Fletcher, Francis, Denton, Wanzek, Wexler, Cirino, Barth, & Romain, 2008; Vaughn & Fuchs, 2003; VanDerHeyden, 2011).

If one considers the original intent of CBM, that is use as a formative assessment approach to inform and influence the instructional behavior of special education teachers, then the focus of the IUA is on the interpretation and use of the scores as a reflection of student performance and progress in reading proficiency, and on the effects (the consequences) of score interpretation and use on teacher instruction and student performance. The evidential basis for score interpretation and use focuses on the technical adequacy of the scores as indicators of performance and progress in reading, and includes questions such as: (1) Do scores reflect general reading proficiency? (2) Do scores increase with improvements in general reading proficiency? (3) If scores for an individual student do not increase, do teachers respond to the scores by making instructional changes for that student? The consequential basis for score interpretation and use focuses on the values and social consequences associated with use of the measures, and includes questions such as: (1) Do instructional changes lead to improved reading proficiency, and is improved reading proficiency a desired social outcome?

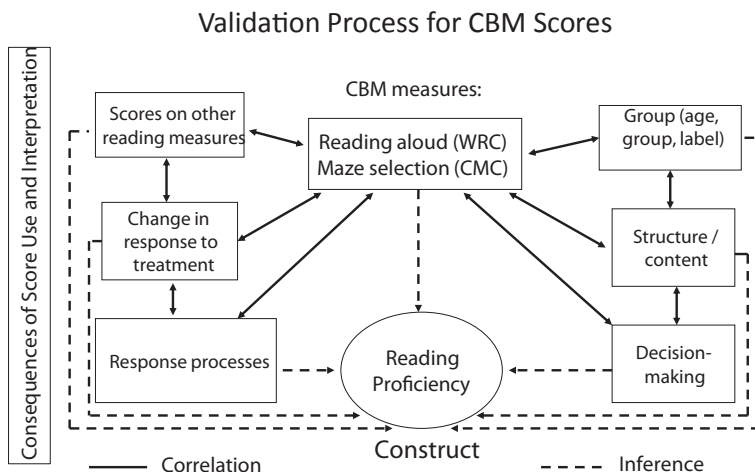
(2) Do the improvements in reading proficiency justify the time, cost, and effort needed to effect such gains? (3) What are the consequences for the school, teachers, and students associated with implementation of a CBM system of progress monitoring for decision-making?

If one considers the use of CBM reading measures within an RTI framework, the focus of the IUA is much different. First, the IUA focuses on the interpretation and use of the CBM scores as potential indicators of the construct LD in reading. Although the evidential basis for score interpretation and use might include the questions mentioned earlier, new questions emerge such as: (1) Do performance and progress on the CBM reading scores adequately differentiate students with and without LD? (2) Do performance and progress on the CBM reading scores adequately identify students in need of different types or intensity of instruction? (3) Does the use of CBM scores within an RTI system lead to better decision-making for identifying students with LD than use of existing procedures? Additional questions also emerge with regard to the consequential basis of score validity, such as: (1) What are the values associated with the label “LD” when it is assigned on the basis of CBM scores within an RTI framework as opposed to when it is assigned using a different approach? (2) What are the social consequences for schools, teachers, students, and families associated with using CBM reading scores to identify students as LD within an RTI system? Serving as a background for each of these questions are the ideologies, values, and beliefs of various stakeholders surrounding the construct of LD.

As can be seen from this brief illustration, the standards and sources of evidence needed to support the validity of score interpretation and use within an RTI framework are different from, and arguably more stringent than, those needed to support the validity of score interpretation and use within a formative assessment framework. The values and consequences associated with the interpretation and use of the scores within a formative evaluation framework relate to teachers modifying instruction to effect higher rates of improvement. The values and consequences associated with interpretation and use of the scores within an RTI framework relate to students receiving a label and being assigned to a particular level or type of instruction. Within both frameworks, the costs of score use and interpretation must be weighed against the benefits to determine the extent to which the scores are “valid enough” to be used to in making each set of decisions.

### ***Validity Is a Unified Concept: Both the Evidential and Consequential Basis of Interpretation and Use of WRC and CMC Need to be Considered***

One needs to simultaneously consider both the evidential and consequential basis of the scores produced by the reading aloud and maze selection measures. The selection of which type of evidence and which consequences to examine depends on the desired interpretation and use of the scores outlined in the IUA, as described in the previous section. To illustrate the unified concept of validity, we focus on the traditional interpretation and use of CBM scores within a formative assessment framework.



**Fig. 13.1** Sample nomological network describing the hypothesized set of relations between scores on common reading assessment (i.e., oral reading fluency and maze) and the construct of proficient reading

**Evidential Basis for WRC and CMC** Recall that there are different types of evidence that support the validity of scores, and that it is important to select which types of evidence are most important for the proposed interpretations and uses (Kane, 2013; Messick 1989a, 1989b). Potential sources of evidence include content, internal structure, response processes, and relations to other variables.

CBM measures are meant to serve as brief indicators of performance and progress in a broad domain; thus, traditionally there has been less interest in and research on content, internal structure, and response processes of the measures, and more on the relations between CBM scores and other variables that represent reading proficiency. However, as will be argued later, examination of the content, internal structure, and response processes might be fruitful and important areas for future research.

An IUA for WRC or CMC is illustrated in Fig. 13.1. The figure resembles a nomological network, as described by Cronbach and Meehl (1955). Cronbach and Meehl (1955) laid the groundwork for a scientific approach to construct validity when they claimed that

“Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses which are a means of confirming or disconfirming the claim” (p. 290).

A nomological network lays out the hypothesized set of relations between the scores on the measurement instrument and the construct.

Figure 13.1 expands upon the concept of nomological networks as described by Cronbach and Meehl (1955) to reflect the view of validity as a unified concept that takes into account both the evidential and consequential basis of score interpretation and use. The solid lines in the figure represent empirical relations, and have arrows

at both ends to represent the bidirectional association among scores on the various measures. The dotted lines represent inferences that are made about the relations between the scores and the construct. The box on the left-hand side of the figure represents the consequences of test use and interpretation, indicating that consequences and evidence interact to form the whole of construct validity.

With regard to the evidential basis of CBM score validity, the figure illustrates that, based on assumptions about the theory underlying measure interpretation and use, there are a set of hypothesized relations that can be tested to examine the validity of scores generated from the CBM reading measures. For example, if the WRC and CMC scores are indicators of general reading proficiency, they should relate in expected ways to selected criterion, such as scores on other measures of reading proficiency. In addition, they should reflect differences in scores by group (e.g., older and younger students, students in higher or lower reading groups, and students with and without LD), and should change in response to effective reading interventions. Their use should produce relevant outcomes, such as better instructional decisions that lead to improved reading performance. Finally, the structure of the measures and response processes used to complete the tasks should fit theoretical rationales underlying the construct of reading proficiency.

In general, there is a fairly large body of research supporting validity claims related to the pattern of relations between the CBM reading scores and scores on other indicators of reading proficiency, and also a fair amount of research demonstrating that scores change in response to interventions (see O'Connor, Gutierrez, Teague, Checca, Sun, & Ho, 2013 and reviews by Marston, 1989; Reschly et al., 2009; Stecker et al., 2005; Wayman et al., 2007). However, there are other sources of evidence that have only infrequently been tapped in the CBM research, but that could provide fruitful areas of research for future study. In addition, there are areas of concern that are in need of additional research.

Sources of evidence that have been examined infrequently are response processes and the structure/content of the reading aloud and maze selection measures. These sources reflect the relation between scores and the theoretical rationales underlying the scores; that is, the meaning of scores within a theoretical framework of reading proficiency (e.g., see Fuchs, Fuchs, Hosp, & Jenkins, 2001). Examples of these types of studies include: (1) a study by Kranzler, Brownell, and Miller (1998), who examined the relative roles of general cognitive ability, speed and efficiency of cognitive processing, and reading aloud in the prediction of reading comprehension; (2) a study by Kendeou and Papadopoulos (2012), who examined the degree to which different cognitive and language skills (e.g., phonological skills, rapid automatized naming (RAN), orthographic skills, and reading fluency) explained unique variance in scores on the maze selection task; and (3) a study by Shinn, Good, Knutson, Tilly, and Collins (1992), who used confirmatory factor analysis to examine the relation of WRC to the reading process from a theoretical perspective for 3rd- and 5th-grade students. Studies such as these serve to increase our conceptual understanding of CBM scores and the relations between the scores and other variables. In addition, such studies can lead to hypotheses regarding how to improve the CBM measures.

More work on the examination of WRC and CMC within cognitive and theoretical frameworks needs to be done (see for example, van den Broek & White, 2012).

Specifically, studies of students' response processes when completing a reading aloud or maze selection task could be conducted. As an example, in a pilot study conducted at Leiden University last year, we (Espin and colleagues) examined the eye-tracking movements of college-age readers during completion of a maze selection task. We were interested in whether readers looked more often and longer at the correct choices than at the incorrect choices. We conjectured that if readers were building a coherent representation of the text (see van den Broek & White, 2012) while completing the maze task, they would jump more often to and focus longer on the correct choice than on the incorrect choices. The results of a controlled eye-tracking study such as the one just described might help us to understand why scores from the maze and reading aloud function differently for older and younger students (see Espin, Wallace, Lembke, Campbell, & Long, 2010; Tichá, Espin, & Wayman, 2009). The increased availability of technology, such as eye-tracking, response-time software, and magnetic resonance imaging (MRI), allows for closer examination of the response processes involved in completion of CBM measures than was possible 30 years ago.

Not only are there new sources of evidence that could be tapped but also there are areas of concern in the CBM research that have not yet sufficiently been addressed. First, with regard to the patterns of relations, more work is needed to examine whether validity results apply equally well to various populations of students, for example, students of different ages, sex, and ethnic/language backgrounds (see Wayman et al., 2007). Some work has been done to examine generalizability of score interpretations and uses to younger (e.g., Dion, Dubé, Roux, Landry, & Bergeron, 2012; Good, Kaminski, Fien, Powell-Smith, & Cummings, 2012) and older students (e.g., Espin & Campbell, 2012), to students of different ethnic and language backgrounds (e.g., Deno & Marston, 2006; Robinson, Robinson, & Blatchley, 2012; Yeo, Ferrington, & Christ, 2011), to different disability groups (e.g., Wallace & Tichá, 2012), and even to students from different countries (e.g., Linan-Thompson, 2012; Shin, 2012), but more work is needed in each of these areas.

Second, a worrisome source of construct-irrelevant variance is the within-individual variability in CBM scores resulting from repeated measurements over time, and the resulting large errors associated with slope estimates (Ardoin & Christ, 2009; Fuchs, 2004; Fuchs & Fuchs, 1992; Ardoin, Christ, Morena, Cormier, & Klingbeil, 2013; Wayman et al., 2007). A related concern is the sensitivity of slope to passage set and passage order effects (Ardoin & Christ, 2009; Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008). Simply put, scores for individuals on "parallel passages" vary a great deal, introducing error into the growth rates produced by CBM measures (Ardoin et al., 2013; Dunn and Eckert, 2002). A reliable and valid slope (rate of growth) for an individual is important if teachers are to make adequate decisions regarding the effects of instruction on student progress. (It is imperative if growth rates are used as a part of RTI decision-making.)

There have been various proposals for ways to deal with within-individual variability, for example, using statistical methods to equate scores (Francis et al., 2008), increasing the number of data points used to calculate slope (Ardoin & Christ, 2009; Christ, 2006), using passage sets that produce less within individual variability (Ardoin & Chris, 2009; Hintze & Christ, 2004), and using generalizability theory to

study sources of error variance (Hintze, Owen, Shapiro, & Daly, 2000). An additional option not often discussed in the literature would be to graph the moving median data rather than raw data.

Another approach that might be used to address the problem of within-individual variability would be to examine the content and structure of the text materials used in CBM. For example, with regard to reading aloud, research in reading comprehension reveals that understanding of a text is influenced by the causal structure of the text (van den Broek & White, 2012). Perhaps passage equivalence could be increased, and variability decreased, if passages were written to reflect a similar causal structure. With regard to maze, consideration of the methods used to select and insert choices into the text might be useful. Maze choices are typically inserted every seventh word and replaced by the correct choice and two distracters, which are within one letter in length of the correct choice (Fuchs & Fuchs, 1992). The correct choice is clearly correct and the incorrect choices clearly incorrect. Little research has been conducted to examine whether the placement or selection of choices influences the validity, reliability, or variability of maze scores (for exceptions see Ketterlin-Geller, McCoy, Twyman, & Tindal, 2006; Parker, Hasbrouck, & Tindal, 1992). Perhaps placement and distracter selection influence passage variability, and could thus be manipulated to reduce variability.

Thus, although there has been a large amount of research addressing issues related to the validity of the scores produced by reading aloud and maze selection measures, as the use of the measures develops and changes, as our knowledge of reading grows and changes, and as our statistical methodologies become more advanced, new questions will arise that can be addressed in new ways. Validity is an ongoing process which brings us to our last point: Validity is a matter of degree.

### ***Validity Is a Matter of Degree***

Validity is a matter of degree. As Messick (1989b) asserts, “validity is an evolving property and validation is a continuing process” (p. 13). Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case possible to guide interpretation and use of test scores (Messick, 1989b). Over time, “the existing evidence becomes enhanced (or contravened) by new findings” (Messick, 1989b, p. 13).

Viewing validation as ongoing process brings us back to the first point, the need for an interpretive/use argument to guide the validation process. Kane (2013) points out that validity changes over time as the interpretations and uses develop, and as new evidence accumulates. If evidence contradicts the arguments laid out in the original IUA, it does not necessarily imply that the scores produced by the measure are invalid; however, it does signal a need to consider potential reasons for the contradictory evidence. Messick (1989b) argues that negative evidence can be interpreted in different ways: The test might not capture the construct very well, the theory underlying the construct might be faulty, the testable hypotheses laid out in the IUA might be faulty, the experimental con-



ditions might not allow for appropriate evaluation of the hypotheses, or a combination of these. Potential responses to contradictory evidence include modifying the measures, modifying the theory underlying the measures, or modifying and testing different hypotheses to explain the contradictory results (Messick, 1989b).

Returning to the example provided earlier—the problem of within-individual variability in scores over repeated measurements and the potential impact this can have on the interpretation and use of CBM reading scores—it is important to discuss and scrutinize the problem (e.g., Ardoin et al., 2013; Wayman et al., 2007). However, it is also important to consider what the implications are for the IUA. Is it possible to modify the measures or the scores in the ways described previously to correct the problem? Does the theory underlying the scores need to be changed? Should scores from the measures be interpreted or used in different ways, and what are the implications associated with such revised interpretations and uses?

Given that research demonstrates that when teachers use individual progress data to inform instructional decisions, performance improves (see review by Stecker & Fuchs, 2000; Stecker et al., 2005), it would seem worthwhile to further examine the issues related to interpretation and use of slope for decision-making. In what ways are the data driving and influencing teachers' decision-making? How much error in slope can be tolerated? (For example, is it the exact slope value that drives decisions or a general pattern of positive or negative growth?) What is the active ingredient leading to student achievement gains: the data, teachers' response to the data, teachers' persistence in striving toward achievement gains, teachers' increased sense of control over the learning trajectories of their students?

All of these questions, and probably more, should be considered in addressing the issue of slope because validity is a matter of degree. It is not possible to say that the WRC and CMC scores are or are not valid—only that the evidence accumulated to date does or does not support particular interpretations and use. If evidence appears that does not support the original interpretations and use, it is incumbent upon the researchers to try and understand these findings and continue the process of validation.

## Conclusion

In conclusion, in this chapter we review current conceptualizations of validity and describe four tenets of validity: validity is a property of a score, not a test instrument; validity is a unified concept; validity is matter of degree; and validation should be guided by interpretation and use arguments. We consider these tenets with respect to scores from two measures that involve fluent reading, reading aloud from text and maze selection, and describe each tenet within a CBM context. We highlight the need for future researchers in the area to be guided by a clear conception of validity with particular regard for clearly stated interpretation and use of the scores.



To conclude, we return to the question posed at the beginning of the chapter: The oral reading fluency measure (ORF):

- A. produces scores that measure reading fluency
- B. produces scores that measure general reading proficiency
- C. produces scores that are indicators of general reading proficiency
- D. should be referred to as a reading aloud measure (RA)
- E. all of the above
- F. none of the above

As we have hopefully illustrated in this chapter, the answer to the question depends on the construct that is being tapped and on the potential interpretations and uses of the scores produced by measures. If we were to answer the question within a CBM context, we would select answer D: Within CBM, an “oral reading fluency” measure is better referred to as a “reading aloud” measure because scores from the measure are used to reflect reading proficiency, not reading fluency. If the question were worded in such a way to ask about the reading aloud measure, then, we would select choice C: The CBM reading aloud measure produces scores that are indicators of general reading proficiency. Note the scores do not measure general reading proficiency. The construct *general reading proficiency* is most closely measured, or represented by, scores on a host of different measures, as illustrated in Fig. 13.1. The CBM scores are intended to be brief indicators of the construct. They are designed to be administered repeatedly to produce scores that, when put on a graph, create a picture of a student’s reading growth over time (Deno, 1985). Through an ongoing and ever-changing process of validation, we can make judgments about whether the existing evidence for and potential consequences of score interpretation and use support our claims about the validity of these CBM scores.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1–15.
- Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review*, 38, 266–283.
- Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding curriculum-based measurement of oral reading fluency (CBM-R) decision rules. *Journal of School Psychology*, 51, 1–18.
- Borsboom, D., & Mellenbergh, G. J. (2004). The concept of validity. *Psychological Bulletin*, 111, 1061–1071. doi:10.1037/0033-295X.111.4.1061.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 546–553.
- Christ, T. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimates of standard error of slope to construct confidence intervals. *School Psychology Review*, 35, 128–133.

- Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justification of test use. *Psychological Methods, 17*, 31–43. doi:10.1037/a0026975.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 292–232.
- Deno, S. L., & Fuchs, L. S. (1987). Developing curriculum-based measurement systems for data-based special education problem-solving. *Focus on Exceptional Children, 19*(8), 1–16.
- Deno, S. L., & Marston, D. (2006). Curriculum-based measurement of oral reading: An indicator of growth in fluency. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 179–203). Newark: International Reading Association.
- Dion, É., Dubé, I., Roux, C., Landry D., & Bergeron, L. (2012). How curriculum-based measures help us to detect word recognition problems in first graders. In C. A. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.), *A measures of success: The influence of curriculum-based measurement on education* (pp. 101–112). Minneapolis: University of Minnesota Press.
- Dunn, E. K., & Eckert, T. L. (2002). Curriculum-based measurement in reading: A comparison of similar versus challenging material. *School Psychology Quarterly, 17*, 24–46.
- Espin, C. A., & Campbell, H. (2012). They're getting older ... but are they getting better? The influence of curriculum-based measurement on programming for secondary-school students with learning disabilities. In C. A. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.), *A measures of success: The influence of curriculum-based measurement on education* (pp. 149–164). Minneapolis: University of Minnesota Press.
- Espin, C. A., & Tindal, G. (1998). Curriculum-based measurement for secondary students. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 214–253). New York: Guilford.
- Espin, C. A., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a progress measurement system in reading for middle-school students: Monitoring progress towards meeting high stakes standards. *Learning Disabilities Research and Practice, 25*, 60–75.
- Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 746*, 315–342.
- Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research and Practice, 18*, 172–186.
- Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*, 188–192.
- Fuchs, D., & Deshler, D. D. (2007). What we need to know about responsiveness to intervention (and shouldn't be afraid to ask). *Learning Disabilities Research and Practice, 22*, 129–136.
- Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*, 45–58.
- Fuchs, L. S., Fuchs, D., Hosp, M., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239–256.
- Fuchs, D., Mock, D., Morgan, P. L., Young, C. L. (2003). Responsiveness-to-Intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research and Practice, 18*, 157–171.
- Good, R. III, Kaminski, R. A., Fien, H., Powell-Smith, K., & Cummings, K. D. (2012). How progress monitoring research contributed to early intervention for and prevention of reading difficulty. In C. A. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.), *A measures of success: The influence of curriculum-based measurement on education* (pp. 113–124). Minneapolis: University of Minnesota Press.

- Hintze, J. M., & Christ, T. J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review*, *33*, 204–217.
- Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly*, *15*, 52–68.
- Hood, S. B. (2009). Validity in psychological testing and scientific realism. *Theory Psychology*, *29*, 451–473. doi:10.1177/0959354309336320.
- Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, *103*, 219–230. doi:10.1007/s11205-011-9843-4.
- Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (Eds.). (2007). *Handbook of response to intervention: The science and practice of assessment and intervention*. New York: Springer.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.
- Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.
- Kendeou, P., & Papadopoulos, T. C. (2012). The use of curriculum-based measurement maze in Greek: A closer look at what it measures. In C. A. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.), *A measures of success: The influence of curriculum-based measurement on education* (pp. 329–339). Minneapolis: University of Minnesota Press.
- Ketterlin-Geller, L. R., McCoy, J. D., Twyman, T., & Tindal, G. (2006). Using a concept maze to assess student understanding of secondary-level content. *Assessment for Effective Intervention*, *31*, 39–50.
- Kranzler, J. H., Brownell, M. T., & Miller, M. D. (1998). The construct validity of curriculum-based measurement of reading: An empirical test of a plausible reival hypothesis. *Journal of School Psychology*, *36*, 399–415.
- Linan-Thompson, S. (2012). Expanding the use of curriculum-based measurement: A look at Nicaragua. In C. A. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.), *A measures of success: The influence of curriculum-based measurement on education* (pp. 321–328). Minneapolis: University of Minnesota Press.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researchers*, *36*, 437–448. doi:10.3102/0013189 × 07311286.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(Monograph Suppl 9), 635–694.
- Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York: Guilford.
- Marston, D., Muyskens, P., Lau, M., & Canter, A. (2003). Problem-solving model for decision making with high-incidence disabilities: The Minneapolis experience. *Learning Disabilities Research & Practice*, *18*, 187–200.
- Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*, 5–11.
- Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.
- Moss, P. A. (2007). Reconstructing validity. *Educational Researcher*, *36*, 470–476. doi:10.3102/0013189 × 07311608.
- O'Connor, R., Gutierrez, G., Teague, K., Checca, C., Kim, J. S., & Ho, T. (2013). Variations in practice reading aloud: Ten versus twenty minutes. *Scientific Studies of Reading*, *17*, 134–162. doi:10.1080/10888438.2011.624566.

- Parker, R., Hasbrouck, J. E., & Tindal, G. (1992). The maze as a classroom-based reading measure: Construction methods, reliability, and validity. *Journal of Special Education, 26*, 195–218.
- Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*, 427–469. doi:10.1016/j.jsp.2009.07.001.
- Reynolds, C. R., & Livingston, R. B. (2012). *Mastering modern psychological testing: Theory and methods*. Boston: Pearson.
- Robinson, S. L., Robinson, M. J., & Blatchley, L. A. (2012). Curriculum-based measurement and English language learners: District-wide academic norms for special education eligibility. In C. A. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.), *A measures of success: The influence of curriculum-based measurement on education* (pp. 187–200). Minneapolis: University of Minnesota Press.
- Shin, J. (2012). Footprints of curriculum-based measurement in South Korea: Past, present, and future. In C. A. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.), *A measures of success: The influence of curriculum-based measurement on education* (pp. 315–320). Minneapolis: University of Minnesota Press.
- Shinn, M. R., Good, R. H. III, Knutson, N., Tilly, W. D. III, & Collins, V. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*, 459–479.
- Speece, D. L., Case, L. P., Molloy, D. E. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research and Practice, 18*, 147–156.
- Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research and Practice, 15*, 128–134.
- Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*, 795–819. doi:10.1002/pits.20113.
- Tichá, R., Espin, C. A., & Wayman, M. M. (2009). Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading aloud and maze selection measures. *Learning Disabilities Research and Practice, 24*, 132–142.
- Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly, 45*, 270–291. doi:10.1598/RRQ.45.3.1.
- van den Broek, P., & White, M. J. (2012). Cognitive processes in reading and the measurement of comprehension. In C. A. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.), *A measures of success: The influence of curriculum-based measurement on education* (pp. 293–306). Minneapolis: University of Minnesota Press.
- VanDerHeyden, A. M. (2011). Technical adequacy of Response to Intervention decisions. *Exceptional Children, 77*, 335–350.
- Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential pitfalls. *Learning Disabilities Research and Practice, 18*, 137–146.
- Vaughn, S., Fletcher, J. M., Francis, D. J., Denton, C. A., Wanzek, J., Wexler, J., Cirino, P. T., Barth, A. E., & Romain, M. A. (2008). Response to intervention with older students with reading difficulties. *Learning and Individual Differences, 18*, 338–345. doi:10.1016/j.lindif.2008.05.001.
- Wallace, T., & Tichá, R. (2012). Extending curriculum-based measurement to assess performance of students with significant cognitive disabilities. In C. A. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.), *A measures of success: The influence of curriculum-based measurement on education* (pp. 211–224). Minneapolis: University of Minnesota Press.
- Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41*, 85–120.

- Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31*, 412–422. doi:10.1177/0741932508327463.
- Yeo, S. (2011). Reliability generalization of curriculum-based measurement reading aloud: A meta-analytic review. *Exceptionality, 19*, 75–93. doi:10.1080/09362835.2011.562094.
- Yeo, S., Fearington, J., & Christ, T. J. (2011). An investigation of gender, income, and special education status bias on curriculum-based measurement slope in reading. *School Psychology Quarterly, 26*, 119–130. doi:10.1037/a0023021.

# Index

## A

- Argument framework 155, 371
- Assessment 4, 10, 22, 26, 29, 42, 44, 52, 70, 157, 161, 166
  - CBM 53, 168, 224, 226, 244
  - fluency and mathematics 78, 99
  - fluency-based 53, 115
  - high-stakes 116
  - modes of 147
  - newly-developed 25
  - of fluent writing 30, 31
  - of letter writing 35, 36
  - ORF 125, 128, 130, 225, 226, 230
  - reading fluency 124
  - sentence-level 40, 41
- Automaticity 4, 5, 12, 29, 49, 50, 187
  - in procedural skill 7, 8, 10
  - theory of 93
- Autoregressive 270, 305, 337, 356

## C

- CBM-writing 31
- Classical test theory (CTT) 145, 146, 150–155, 167, 174, 176, 339
- Cluster analysis 313, 315
- Conditional item response theory (CIRT) 166, 170, 171, 173, 176, 177, 181
  - analysis of 183
- Curriculum-based measurement (CBM) 4, 23, 30, 53, 91, 95, 107, 114, 143, 166, 318, 371
  - fluency-based 109
  - in LSN 110
  - multitiered 54
  - of ORF 124, 130
  - principles of 26, 27

## D

- Data-based decision making 91, 99, 125
  - basic components of 100, 101
- Decision-making 325, 326
- Diagnostic accuracy 212

## E

- Educational assessment 146, 171
- Equipercentile equating 236, 238, 243

## F

- Factor analysis 330
- Fluency 1, 70, 71, 94
  - an indicator 149, 150
  - based measures 80–83
    - assessing number sense 80–82
    - in the middle school grades 83
  - definition of 5
  - in language proficiency 1–3
  - in reading 4, 5
  - mathematics interventions with 76, 77
  - measures of 166, 167
  - oral reading 224–226
  - passage reading 233
  - recommendations for, study of 11–14
- Form effects 167, 223, 226, 239

## G

- Generalizability theory (GT) 145, 146, 152
- Growth modeling 271, 272, 280, 291, 297, 300, 305, 362

## I

- Individual growth curves 337
- Instruction 50–52
- Item response theory (IRT) 153

**L**

Language proficiency 1, 2, 6, 12, 319  
 Latent change 305, 334, 356  
 Latent class analysis 313  
 Latent growth 294, 334, 335, 338, 341, 343, 344, 346, 348  
 Latent profile analysis (LPA) 311, 317  
 Latent variable equating 232, 247, 248, 254  
 Linear equating 234, 236, 243

**M**

Mathematics 95, 99, 100, 116  
   assessment 78, 99  
   fluency 5, 6, 67, 69, 72, 73, 75, 76, 78  
   instruction 72, 73, 76, 77, 83  
 Measurement 168, 169  
   directness of 146, 147  
 Mixture model 330

**N**

Number sense 72, 77, 79, 111  
   domain of 84  
   fluency-based measures assessing 80–82

**O**

Oral reading fluency (ORF) 3, 102, 109, 124, 197, 198, 202, 204, 213, 224, 273, 338, 345  
   CBM 115  
   definition of 4  
   of measuring 12

**P**

Progress monitoring 43, 46, 54, 58, 79, 101, 102, 237  
   follow-up 108  
   system of 95, 374  
 Psychometrics 168

**R**

Reliability 35, 37, 46, 168  
   alternate-form 45, 79, 151  
   in CTT 154

  inter-rater 130  
   model-based, comparison of 179, 181  
   of CBM 95, 226  
   test-retest 38, 43, 131  
 Repeated measures 73, 269, 274, 279  
   of ORF 297  
 Response to intervention (RTI) 39, 54, 101, 125  
 ROC curve 195–197, 199, 200, 206, 211, 215, 216

**S**

Science, technology, engineering, and mathematics (STEM) 67, 68  
 Screening 54, 215  
   CBM 102, 106, 108, 109, 116  
   short-term 58  
 Signal detection theory 189, 192, 199, 200, 205, 207  
   recommendations for, application of 214  
 Speeded assessment 147  
 Structural equation modeling (SEM) 48, 270, 280, 281, 361

**T**

Test utility 197, 224

**U**

Universal screening 54, 101

**V**

Validity 35, 41, 42, 156, 366  
   consequential, basis of 369, 370  
   evidential, basis of 367–369  
   matter of degree 371, 378, 379  
   measures involving fluent reading, of scores 372  
   of WRC 372  
   unified concept, all of 367

**W**

Writing fluency 23, 36, 49, 51, 92  
 Written expression 27, 43, 99