Methodology of Educational Measurement and Assessment

Alina A. von Davier
Mengxiao Zhu
Patrick C. Kyllonen *Editors*

Innovative Assessment of Collaboration

Springer

Methodology of Educational Measurement and Assessment

Series editors

For avoidance of conflicts of interest, M. von Davier has not been involved in the decision making process for this edited volume.

This new book series collates key contributions to a fast-developing field of education research. It is an international forum for theoretical and empirical studies exploring new and existing methods of collecting, analyzing, and reporting data from educational measurements and assessments. Covering a high-profile topic from multiple viewpoints, it aims to foster a broader understanding of fresh developments as innovative software tools and new concepts such as competency models and skills diagnosis continue to gain traction in educational institutions around the world. Methodology of Educational Measurement and Assessment offers readers reliable critical evaluations, reviews and comparisons of existing methodologies alongside authoritative analysis and commentary on new and emerging approaches. It will showcase empirical research on applications, examine issues such as reliability, validity, and comparability, and help keep readers up to speed on developments in statistical modeling approaches. The fully peer-reviewed publications in the series cover measurement and assessment at all levels of education and feature work by academics and education professionals from around the world. Providing an authoritative central clearing-house for research in a core sector in education, the series forms a major contribution to the international literature.

More information about this series at http://www.springer.com/series/13206

Alina A. von Davier · Mengxiao Zhu
Patrick C. Kyllonen
Editors

Innovative Assessment of Collaboration

Springer

Editors
Alina A. von Davier
ACT
Iowa City, IA
USA

Patrick C. Kyllonen
Research and Development Division
Educational Testing Service
Princeton, NJ
USA

Mengxiao Zhu
Research and Development Division
Educational Testing Service
Princeton, NJ
USA

ISSN 2367-170X ISSN 2367-1718 (electronic)
Methodology of Educational Measurement and Assessment
ISBN 978-3-319-33259-8 ISBN 978-3-319-33261-1 (eBook)
DOI 10.1007/978-3-319-33261-1

Library of Congress Control Number: 2016961330

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Part I Framework and Methods

Part II Modeling and Analysis

Contributors

Vincent Aleven Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Raquel Asencio Purdue University, West Lafayette, IN, USA

Yoav Bergner Educational Testing Service, Princeton, NJ, USA; New York University, New York, NY, USA

Paul B. Borysewicz Educational Testing Service, Princeton, NJ, USA

Sy-Miin Chow Pennsylvania State University, State College, USA

Danielle Clewley University of Memphis, Memphis, TN, USA

Jeffrey F. Cohn University of Pittsburgh, Pittsburgh, USA

Noshir Contractor Northwestern University, Evanston, IL, USA

Alina A. von Davier ACT, Iowa City, IA, USA

Matthias von Davier Educational Testing Service, Princeton, NJ, USA

Leslie A. DeChurch Northwestern University, Evanston, USA

Nia Dowell University of Memphis, Memphis, TN, USA

Jennifer Feitosa City University of New York, New York, USA

Oliver Ferschke Language Technologies Institute and Human-Computer Interaction Institution, Carnegie Mellon University, Pittsburgh, PA, USA

Stephen M. Fiore University of Central Florida, Orlando, FL, USA

Trysha Galloway The Learning Chameleon Inc, Culver City, CA, USA

Michele Gelfand University of Maryland, College Park, USA

Arthur C. Graesser University of Memphis, Memphis, TN, USA

Samuel Greiff Computer-Based Assessment Research Group, University of Luxembourg, Luxembourg City, Luxembourg

Patrick Griffin Melbourne Graduate School of Education, Parkville, Australia

Markku T. Hakkinen Educational Testing Service, Princeton, NJ, USA

Peter F. Halpin New York University, New York, NY, USA

Jiangang Hao Educational Testing Service, Princeton, NJ, USA

Qiwei He Educational Testing Service, Princeton, NJ, USA

Iris Howley Language Technologies Institute and Human-Computer Interaction Institution, Carnegie Mellon University, Pittsburgh, PA, USA

Katelynn A. Kapalo University of Central Florida, Orlando, FL, USA

Saad M. Khan Educational Testing Service, Princeton, NJ, USA

Patrick C. Kyllonen Research and Development Division, Educational Testing Service, Princeton, USA

Cynthia Lamb URS Federal Technical Services Inc, Philadelphia, PA, USA

Jerry Lamb Naval Submarine Medical Research Laboratory, Groton, CT, USA

Lei Liu Educational Testing Service, Princeton, NJ, USA

Daniel S. Messinger University of Miami, Coral Gables, USA

Amy Ogan Carnegie Mellon University, Pittsburgh, PA, USA

Jennifer K. Olsen Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Lu Ou Pennsylvania State University, State College, USA

Denise L. Reyes Rice University, Houston, USA

Carolyn Penstein Rosé Language Technologies Institute and Human-Computer Interaction Institution, Carnegie Mellon University, Pittsburgh, PA, USA

Nikol Rummel Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA; Institute of Educational Research, Ruhr-Universität Bochum, Bochum, Germany

Eduardo Salas Rice University, Houston, USA

Aaron Schecter Northwestern University, Evanston, IL, USA

C. Shawn Burke University of Central Florida, Orlando, USA

Ron Steed UpScope Consulting Group, Mysti, CT, USA

Eric W. Steinhauer Educational Testing Service, Princeton, NJ, USA

Ronald Stevens UCLA School of Medicine, Los Angeles, CA, USA; The Learning Chameleon Inc, Culver City, CA, USA

Tracy M. Sweet University of Maryland, College Park, MD, USA

Erin Walker Arizona State University, Tempe, AZ, USA

Miaomiao Wen Language Technologies Institute and Human-Computer Interaction Institution, Carnegie Mellon University, Pittsburgh, PA, USA

Jason J.G. White Educational Testing Service, Princeton, NJ, USA

Amanda L. Woods Rice University, Houston, USA

Diyi Yang Language Technologies Institute and Human-Computer Interaction Institution, Carnegie Mellon University, Pittsburgh, PA, USA

Mengxiao Zhu Research and Development Division, Educational Testing Service, Princeton, USA

Prologue

Building the Foundation for Innovations in Assessment: Interdisciplinary Measurement of Collaboration and Teamwork[1]

This volume is intended to encourage and inspire researchers from across many disciplines to begin or grow efforts focused on overcoming some of the scientific barriers to achieving major innovations in assessment. As a vehicle for discussion of these barriers, scientists from the US Army Research Institute (ARI) and Educational Testing Service have focused the challenge of how to build assessments of collaborative and teamwork skills and performance. A working meeting was held in November 2014 to initiate this discussion, and this volume follows up on that meeting to broaden this discussion into the full interdisciplinary research community focused on assessment and measurement of collaboration and teamwork. The chapters in this interdisciplinary volume present a variety of perspectives on approaches to measuring collaboration, communication, and interactions. We hope this discussion will serve as a foundation for the future development of assessments—for applications in both educational and organizational/workforce settings. We are particularly optimistic about the possibility of transferring findings from several disciplinary perspectives on the measurement of collaboration and teamwork to educational settings.

Considerable research has been accomplished and lessons learned from studying teams in organizational settings, as several of the chapters in this volume attest. The chapters by Salas, Reyes, and Wood; Asencio and DeChurch; and Fiore and Kapalo

[1] This work was conducted while Alina A. von Davier was employed with Educational Testing Service.

provide a view into the depth of research on measurement of teamwork and collaboration in organizations. However, comparatively little has been done on assessing collaboration within educational settings, with some notable exceptions. The chapters by Olsen, Aleven, and Rummel; He, M. von Davier, Greiff, Steinhauer, and Borysewicz; and Griffin illuminate some of these efforts, and particularly highlight some of the work initiated under the Assessment and Teaching of 21st Century Skills (ATC21S). That work was continuing through the Collaborative Assessment Alliance (CAA), under the leadership of the late Greg Butler, who attended the *Innovations in Collaboration* working meeting, before his untimely passing.

We believe the objective of initiating this discussion, through the working meeting and subsequently through this volume, has been met and has great hopes this volume will provide the basis for growing an emphasis on identifying and overcoming the scientific barriers to substantial innovation in measurement and assessment. Chapters from experts approaching collaboration from the standpoint of several areas of research were assembled in this volume. The realm of applications collected in this volume was quite rich, ranging from intelligent tutoring systems and simulation-based achievement tests to submarine/navy teams, basketball teams, and mother–child interactions. Technologies included dual eye tracking, sociometric badges, animation dialogues called *trialogues*, massive open online courses (MOOC) and Wikipedia collaborations, assistive technologies, and electroencephalography (EEG). Methodologies were equally diverse, ranging from multilevel modeling to social network analysis and relational events modeling, point processes, hidden Markov models, and dynamic systems modeling techniques. The research and concepts presented here represent an extraordinary range of potential approaches to assessing, measuring, and ultimately modeling collaboration and teamwork. While many, if not all, of these approaches provide insight into potential new methods for assessment, they also help illuminate some of the challenges we all face in revolutionizing measurement. We hope that the breadth of applications, technologies, and methods presented here will inspire the development of the next generation of assessments of collaborative skills and teamwork.

As briefly noted earlier, this volume is the most recent step in a chain of events focused on moving the science of measurement forward toward major innovations in measurement theory and assessment techniques. In 2012, ARI engaged the National Research Council (NRC) to execute a consensus study entitled *Measuring Human Capabilities: Performance Potential of Individuals and Collectives* (NRC, 2015). The initial phase of this consensus study was an NRC-hosted workshop, which produced the report *New Directions in Assessing Performance Potential of Individuals and Groups* (NRC, 2013). This chapter's authors all played a role in that workshop—Gerald Goodwin was the study sponsor, Patrick Kyllonen was a member of the study committee, and Alina von Davier was a keynote speaker. The *New Directions* workshop was planned and structured to identify and discuss good ideas that were at the forefront of innovations in measurement of individuals and collectives. The consensus study further explored those ideas as well as others and honed down to a final list of strong recommendations, which are summarized in the study report *Measuring Human Capabilities: An Agenda for*

Basic Research on the Assessment of Individual and Group Performance Potential for Military Accession (NRC, 2015). However, from the *New Directions* workshop, scientists at ETS proposed to begin exploring how to identify and accelerate innovations in assessment with a particular emphasis on collaboration. Under a research grant from ARI, ETS assembled a working meeting titled *Innovations in Collaboration*, held in November 2014 in Washington, DC. From the workshops and working meeting, we concluded that to learn more about communication, collaboration, and how to build assessments for these concepts, we need to work together across disciplines and learn from each other's perspectives in order to accelerate the research around this new type of assessment. Therefore, this interdisciplinary conversation is the primary goal for this volume.

Historical Perspective

As previously noted, ARI began this focus on innovations in assessment in 2012. In order to understand why this focus was brought up, it is helpful to step back and gain some perspective on measurement and assessment, particularly within the military, over the last century and more. The first half of the twentieth century was dominated by a measurement paradigm commonly known as classical test theory (see Novick, 1966 for an integrated review). The origins of military entrance testing also go back to this era. In 1917, ARI's predecessor organization created the Army Alpha and Army Beta to improve the accession and classification of military recruits during World War I and both of these tests were firmly grounded in the classical test theory paradigm. Over the following decades, ARI's direct ancestors developed a series of entrance and classification tests, all continuing to be grounded in the classical test theory paradigm. While the paradigm is quite functional—indeed, it is still used as the basis for a wide variety of psychological tests—it also has several weaknesses.

In the 1950s, the science of measurement began a major paradigm shift. One of the breakthroughs in the development of selection and classification testing was item response theory (IRT), which is currently the dominant measurement paradigm. The original work on IRT began in the 1950s. Seminal papers for IRT were written by Fred Lord of ETS (1952, 1953a, b). That work led to conceptual demonstrations and discussion of how we could apply IRT to generate adaptive tests in the 1970s (Lord, 1974; Weiss, 1976). Breakthroughs in computing and microprocessor technology led to being able to put IRT models into play in the form of computer-adaptive tests in the 1970s. The Department of Defense—including ARI, as well as sister laboratories in the Air Force and Navy—began work to develop a new military entrance test that could be implemented as a computer-adaptive test and build from the IRT framework instead of classical test theory. This new military entrance test was called the Armed Services Vocational Aptitude Battery (ASVAB). ETS was engaged in a similar endeavor at that time to put into practice the first computer-adaptive test for the Graduate Record Examination (GRE), the Test of English as a Foreign Language (TOEFL), and the

General Management Admission Test (GMAT). These parallel development efforts during the 1970s and 1980s led to computer-adaptive versions of the ASVAB as well as the ETS tests going operational in 1990, with a computer-adaptive prescreening version for ASVAB that went operational in 1988 (CAT GRE went operational in 1993).

The next breakthrough was the development of a computer-adaptive personality test for the Army. Although many studies looking at personality as it relates to job performance were conducted over the years (including the origin of the five-factor model of personality within the Department of Defense, DoD; Tupes & Christal, 1961), the validity of these tests, and particularly the incremental validity beyond tests of general mental ability, was consistently quite low when put into operational use. Quite simply, there were persistent problems in the susceptibility of personality measures to faking and intentional response distortion. However, the development of adaptive IRT-forced-choice methodologies (Stark, Chernyshenko, Drasgow, & White, 2012) made it possible to build a personality assessment that was resistant to faking. Ultimately, this led to the development and operational implementation of a computer-adaptive personality test called the Tailored Adaptive Personality Assessment System (TAPAS; Drasgow et al., 2012).

Looking Forward

As ARI's Basic Research Office was re-evaluating its program emphasis on selection-related research between 2008 and 2011, it became clear that a renewed emphasis on accelerating developments in psychometric theory was needed. Looking backward, the original theoretical work on IRT was done in the 1950s. While the conceptual development of IRT matured through the 1960s, it was not until the advent of sufficiently powerful computers that IRT was able to be put into practical use for adaptive testing. As such, it took nearly 35 years to reach the first computer-adaptive test, and another 20 years of incremental advances within IRT to reach a second major transition to measure personality. While there is clearly additional work to be done within the IRT paradigm that will continue to produce innovations in measurement in the future, the scientists at ARI noted the very long timescale to evolve from serious original psychometric theory to the practical application of that work. As such, ARI's basic research scientists began this emphasis on finding, inspiring, and accelerating deep original psychometric theory that will overcome many of the limitations of IRT and produce the next generation of assessments and tests.

New Constructs, New Methods

A principal motivation for ARI's interest in this area is the need to anticipate and begin encouraging and supporting the science that is required before the next major generation of military entrance tests can be developed and implemented to replace

ASVAB 20–25 years in the future. Contemplating the future of assessment entailed introspection and conversation with experts in military personnel testing, educational assessment experts, psychometricians, and others; this resulted in the ARI team understanding that there are several critical issues that cannot be addressed very well or at all within the current testing paradigm. For example, ARI has been a significant proponent of and investor in research on situational judgment testing, going back almost two decades. But a persistent and troubling finding with these tests, as well as other types of performance-based tests, has been that multiple constructs are inevitably called upon in performance and it is difficult or impossible to assess specific individual attributes separate from performance. This leads to items that measure several distinct constructs and tests that largely reflect the broad categories of performance rather than the individual difference constructs intended to be assessed. The assumptions within the current underlying psychometric theory do not allow us to deal with this situation very neatly.

Assessment of social and interpersonal skills is another area that often has been suggested by findings and survey results; however, attempts to measure these skills invariably also run into the same issues of complex variance that has been difficult if not impossible to disentangle. A key issue with assessing social and interpersonal skills is that there are invariably multiple component skills being assessed together. Moreover, social and interpersonal interactions are dynamic and not very amenable to assessments through static methods. Being able to assess multiple constructs simultaneously in a dynamic way is challenging with most of our current psychometric tests developed within a (unidimensional) psychometric theory that is focused on measuring one construct at a time. We design items that are ideally suited or ideally would tap into a single construct and we aim to have a pure item assessment. But what happens when we have items or performance sets in which you have multiple constructs contributing simultaneously and dynamically to the performance or response? We know that there are multidimensional psychometric models, even though they are not yet in operational use with large-scale assessments. But what about dynamic models, in which the constructs employed change over time?

New Environment

Technological advances in recent years have made it possible to use computers to capture rich data about complex performance (e.g., the interactions of individuals). Assessment developers have sought to leverage this capability to better understand the processes test takers employ to reach their final answers. Capturing interactions, at least on a large scale, had been impossible without the medium of the virtual environment. Now that virtual media are available for (educational) assessment, advances to assessment are now possible.

A good assessment allows valid inferences about the degree to which a test taker possesses the knowledge, skills, and abilities covered by the assessment. Traditional assessment performance, however, does not always match actual performance in academic or workforce situations, and part of this dissimilarity may be linked to the dissimilarity in the context of the traditional assessment and the context in which knowledge is expected to be applied.

In order to assess cognition from outcome data alone, we must assume that the final answer a test taker provides is in some way indicative of the underlying thought process that produced it. Even when responding to an explicit question, those cognitive processes can vary from very careful reasoning based solely on the test taker's content knowledge to seemingly random responses (guesses) based upon unsystematic or arbitrary choices. Such variation in underlying cognitive processes is not necessarily reflected in the correctness of a test taker's responses, many of which might be dichotomous. Cognitive diagnostic models have been employed to identify relevant factors that contribute to a correct or incorrect response and to identify students' misconceptions using traditional assessments (see Katz, Martinez, Sheehan, & Tatsuoka, 1998; von Davier, 2005). Often these attributes are highly correlated and thus difficult to accurately estimate from the data. Hence, despite these efforts, it is still challenging to provide actionable feedback to students based on traditional items.

Developing assessments in a virtual environment has several attractive qualities that may be used to provide adequate feedback and enhance learning. An organic link might be constructed among teaching, learning, and assessment, and a natural environment can be provided for (virtual) collaboration among test takers, either working in person or in remote teams (von Davier & Mislevy, 2016).

Perhaps most importantly, the intersections of assessment, cognition, and learning in a computerized assessment environment allow us to identify the strategies test takers employ and thus to examine their problem-solving processes. Process data can be recorded and can be used to effectively reconstruct a test taker's actions during the assessment, allowing inferences about aspects of a test taker's cognition based on those actions taken during the assessment. Those process data can be used to analyze the behaviors associated with final responses, in turn allowing us to form actionable hypotheses about how and why test takers provided the responses we see. Obviously, collaborative assessments in a computerized environment would result in rich metadata collected in sophisticatedly designed log files (see Hao, Liu, A. von Davier, and Kyllonen; Bergner, Walker, and Ogan; Halpin and A. von Davier; and Zhu in this volume). Importantly, the type of process data described here might be very valuable to overcoming the challenges inherent in disentangling complex performance described earlier.

Educational Testing

Similarly, at ETS, there has been considerable interest in moving beyond the traditional areas of assessing curricular skills, reasoning ability, mathematics achievement, and English language skills. The establishment of the Center for

Academic and Workforce Readiness and Success, headed by Patrick Kyllonen, and the Computational Psychometrics Research Center, headed by Alina von Davier, represents a commitment by the organization to explore new constructs, new measures, and new psychometrics. This convergence of interests has been central to the emergence of this volume.

ETS has also committed to collaboration as an important new construct based on surveys of employers and educators and the growing importance of collaboration, communication, and social skills as critical skills for the 21st century. ETS currently has numerous projects concerned with exploring measurement and modeling approaches to these skills and examining the evidence for the validity of such measures for the purposes of admissions, program evaluation, formative assessment, and student learning outcomes assessment. For example, ETS developed the collaborative problem-solving tasks in PISA 2015; this volume includes a chapter that describes the innovative collaborative tasks that have been implemented in more than 60 languages and delivered online. Over the past three years, ETS has brought in new scientists with varied backgrounds—several of whom have made contributions to this volume—to tackle the challenges associated with measuring these more complex skills, reflecting a commitment and a long-term investment to improve the way we define, measure, and operationalize collaborative problem-solving. The organization views addressing this challenge as critical for higher education, for K-12, and for the workforce.

We did not have to start from scratch. Within education, there exists a model of collaborative problem-solving that we can build on. The work of the ATC21S consortium, documented in the chapter by Patrick Griffin, describes the overarching framework, processes, skills, and learning progressions that can be measured with an existing, operational system of collaborative problem-solving. We see this as an important framework to expand on. It encourages systematically thinking about developmental progressions, or learning progressions, or novice-to-expert rubrics or stages that provide a new construct-focused measurement context that goes beyond traditional criterion-referenced or norm-referenced systems. This provides a broader conceptual notion of collaboration or coordination and how it develops to expertise. Griffin's work was the pioneering effort for collaborative work education, and when people think about measuring collaborative problem-solving they think of the kinds of tasks that he invented, with problem statements, chat boxes, and online environments.

Next Steps

We intended this volume to inspire renewed enthusiasm for interdisciplinary approaches to addressing the problem of assessing difficult-to-measure attributes and skills, such as collaboration. We emphasize the potential of blended disciplines as in computational psychometrics in addressing the challenges around the measurement of hard-to-measure constructs (von Davier, in press). First and foremost, we hope that the chapters contained within this volume stimulate interest in moving measurement theory forward. The emphasis on assessment of collaboration is an able vehicle for us to do that. Collaboration is a complex performance involving multiple people, both social and interpersonal skills, and complex constructs that require identifying interesting and accurate methods of assessment. We hope this volume provides the inspiration and a framework to move measurement theory forward, to develop new approaches for assessing complex constructs such as social and interpersonal skills, and to identify approaches to measure these complex constructs. This volume hopefully will also inspire researchers at a wide variety of organizations to embrace some of these ideas and help move the science of measurement forward into the future.

Gerald F. Goodwin
Army Research Institute
Alexandria, VA
USA

Patrick C. Kyllonen
Educational Testing Service
Princeton, NJ
USA

Alina A. von Davier
ACT
Iowa City, IA
USA

References

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support army personnel selection and classification decisions* (U.S. ARI Technical Report No. 1311). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Katz, I. R., Martinez, M. E., Sheehan, K. M., & Tatsuoka, K. K. (1998). Extending the rule space methodology to a semantically-rich domain: Diagnostic assessment in architecture. *Journal of Educational and Behavioral Statistics*, *23*, 254–278.

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Richmond, VA: Psychometric Corporation.

Lord, F. M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, *18*, 57–75.

Lord, F. M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, *13*, 517–549.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*, 247–264.

National Research Council. (2013). *New Directions in assessing performance potential of individuals and groups: Workshop summary*. R. Pool, Rapporteur. (Committee on Measuring Human Capabilities: Performance Potential of Individuals and Collectives, Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education). Washington, DC: The National Academies Press.

National Research Council. (2015). Measuring human capabilities: An agenda for basic research on the assessment of individual and group performance potential for military accession. (Committee on Measuring Human Capabilities: Performance Potential of Individuals and Collectives, Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education). Washington, DC: The National Academies Press.

Novick, M. R. (1966). The axions and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*, 1–18.

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods*, *15*(3), 1–25. doi: 10.1177/1094428112444611

Tupes, E. C., & Christal, R. E. (1961). *Recurrent personality factors based on trait ratings* (USAF ASD Technical Report No. 61-97). Lackland Air Force Base, TX: U.S. Air Force.

von Davier, A. A. (n.d.). Computational psychometrics in support of collaborative assessments. *Journal of Educational Measurement* (in press).

von Davier, M. (2005). A general diagnostic model applied to language testing data. (Research Report No. 05-16). Princeton, NJ: Educational Testing Service). doi: 10.1002/j.2333-8504.2005.tb01993.x

von Davier, A. A., & Mislevy, R. J. (2016). Design and modeling frameworks for 21st century: Simulations and game-based assessments. In C. Wells & M. Falkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 239–256). New York, NY: Guilford.

Weiss, D. J. (1976). Adaptive testing research at Minnesota: Overview, recent results, and future directions. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (pp. 24–35). Washington, DC: U.S. Civil Service Commission.

Chapter 1
Introduction: Innovative Assessment of Collaboration

Patrick C. Kyllonen, Mengxiao Zhu, and Alina A. von Davier

Abstract In this introductory chapter we provide the context for this edited volume, describe the recent research interests around developing collaborative assessments around the world, and synthesize the major research results from the literature from different fields. The purpose of this edited volume was to bring together researchers from diverse disciplines—educational psychology, organizational psychology, learning sciences, assessment design, communications, human-computer interaction, computer science, engineering and applied science, psychometrics—who shared a research interest in examining learners and workers engaged in collaborative activity. This chapter concludes with an emphasis on how each chapter contributes to the research agenda around the measurement research questions, from how to define the constructs to how to model the data from collaborative interactions.

Keywords Collaborative problem solving · Assessment · Testing · Educational psychology · Organizational psychology · Learning sciences · Assessment design · Communications · Human-computer interaction · Computer science · Engineering and applied science · Psychometrics

This work was conducted while Alina A. von Davier was employed with Educational Testing Service.

P.C. Kyllonen (✉) · M. Zhu
Research and Development Division, Educational Testing Service,
Princeton, USA
e-mail: pkyllonen@ets.org

M. Zhu
e-mail: mzhu@ets.org

A.A. von Davier
ACT, Iowa city, USA
e-mail: alina.vondavier@act.org

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_1

1.1 Background

Several employer surveys over the past few years attest to the importance of teamwork, collaboration, and communication skills. For example, a recent national survey of college recruiting professionals in government, manufacturing, service, retail, and transportation (National Association of Colleges and Employers, 2014) found that the "ability to work in a team structure" was rated highest in importance for hiring potential employees of the skills considered, which included obviously important ones such as "make decisions and solve problems," "plan, organize, and prioritize work," and "analyze quantitative data." The importance of teamwork skills echoed findings from a similar survey conducted a few years previously by the Conference Board and others (Casner-Lotto & Barrington, 2006) which found that "teamwork/collaboration" and "oral communication" were two of the three applied skills (the other being "professionalism/work ethic") rated most important by employers.

We also conducted a study at ETS based on an analysis of the Department of Labor's Occupational Network (O*NET) database (Burrus et al., 2013), and found that a teamwork factor was rated third in importance behind problem solving and reasoning ability, but ahead of 12 other factors, including achievement, innovation, and information technology literacy. These survey findings are consistent with the competencies employers evaluate their workforce on, such as the Lominger competencies (Korn/Ferry International, 2014–2016) "comfort around management," "developing others," "directing others," "interpersonal savvy," "listening," "peer relationships," "sizing up people," "building effective teams," and "understanding others." The survey findings are also consistent with recent work supporting a view of the growing importance of social skills in the labor market based on greater relative growth of jobs with high social skills requirements and greater wage growth for jobs that require high cognitive and social skills in combination (Deming, 2015; see also, Weinberger, 2014). The growing recognition of the importance of teams and social skills in the workforce has made its way into the popular press, as demonstrated by a recent New York Times magazine article on the topic (Duhigg, 2016).

Schools also are beginning to focus attention on teaching collaboration. For example, the National Research Council (2012) reviewed evidence showing the importance of teamwork and collaboration in schools, and pointed out that the Common Core State Standards emphasizes "collaboration and listening with care to understand and evaluate others' utterances" in the English Language Arts standards (p. 114), and collaboration/teamwork in the mathematics standards (p. 123). This new found emphasis is reflected in the presence of a collaborative problem solving assessment in the 2015 Program for International Student Assessment (PISA) (Organisation for Economic Cooperation and Development, 2013), and in a recent "Innovations Symposium" by the U.S. Department of Education focusing on collaborative problem solving (National Center for Education Statistics, 2014).

1.2 Assessment

Given the apparent growing recognition of the importance of social and collaborative skills why has there been no concomitant improvement in our sophistication for measuring them? It seems that even today the most common approach for measuring teamwork and social skills is to ask people to rate themselves (or others) on a 5-point agreement scale with statements such as “I work well with others,” or “I am described by others as a good ‘team player’.” Such methods for describing oneself or others are extremely useful, serve as the basis for personality psychology (John, Naumann, & Soto, 2008), and have resulted in taxonomies providing robust, cross-situational, and cross-cultural dimensions of behavior, such as the five factor model (John, 1990), and its relatives (Paunonen & Ashton, 2001). And indeed several of the Big 5 dimensions —agreeableness and extroversion—and more fine-grained facet dimensions—social dominance, sociability, warmth, generosity, and cooperation and trust—can be appropriately thought of as components of collaborative skill (Drasgow et al., 2012). It is not surprising that much current research on collaboration and collaborative problem solving particularly with adults in organizations relies on the traditional Likert-scale measure of collaboration (see especially chapters by Salas, Reyes, & Woods, Chap. 2; Acencio & DeChurch, Chap. 3).

However, one of our hopes in assembling this volume was to generate interest in new, innovative approaches to measuring collaboration and collaborative skill. Even self-reports can be improved upon. Anchoring vignettes (King, Murray, Salomon, & Tandon, 2004), in which self-ratings are adjusted by how respondents rate hypothetical others have been proposed for increasing cross-cultural comparability in educational assessment (Kyllonen & Bertling, 2014). Forced-choice methods, in which respondents are asked to choose a statement that best describes them, rather than to rate their agreement with a statement on a five-point scale, have been shown to increase both outcome predictions in school and in the workforce (Salgado & Tauriz, 2014), and cross-cultural comparability (Bartram, 2013). Ratings by others compared to self-ratings have been shown to be more reliable and provide better predictions of future performance (Connelley & Ones, 2010; Oh, Wang, & Mount, 2011). Standardized peer ratings in the form of behaviorally anchored rating scales (BARS) and behavioral observation scales (BOS) are commonly used in organizational settings (see Salas et al., Chap. 2; and Acencio & DeChurch, Chap. 3).

Situational judgment tests, which have been described as “low fidelity simulations” (Motowidlo, Dunnette, & Carter, 1990), also are a promising measurement methodology (Weekley, Ployhart, & Harold, 2004; Whetzel & McDaniel, 2009). An example to measure teamwork and collaboration (Wang, MacCann, Zhuang, Liu, & Roberts, 2009, p. 114, italics added) is the following.

> You are part of a study group that has been assigned a large presentation for class. As you are all dividing up the workload, it becomes clear that both you and another member of the group are interested in researching the same aspect of the topic. Your colleague already has a great deal of experience in this area, but you have been extremely excited about working

on this part of the project for several months. Rate the following approach to dealing with this situation:

(a) Flip a coin to determine who gets to work on that particular aspect of the assignment;
(b) Insist that, for the good of the group, you should work on that aspect of the assignment because your interest in the area means you will do a particularly good job;
(c) Compromise your preferences for the good of the group and allow your friend to work on that aspect of the assignment;
(d) Suggest to the other group member that you both share the research for that aspect of the assignment and also share the research on another less-desirable topic.

An attractive feature of situational judgment tests is that they can measure subtle qualities of judgment and even ones on which there is not necessarily consensus on the best course of action (Zu & Kyllonen, 2012), making them ideally suited for measuring qualities such as teamwork and collaboration. (In this example, no key was provided by the authors, but they stated that the key was decided by “a panel of three assessment specialists in educational and psychological testing” (Wang et al., 2009, p. 29); a sum score was computed as the number of times an examinee’s choice matched the expert key).

More recently there has been considerable interest in measuring collaboration directly through performance on collaborative games and simulations. In this volume, chapters by Griffin (Chap. 8) and Hao, Liu, von Davier, and Kyllonen (Chap. 9) discuss assessment in two (or more)-person interactive games and simulations. Graesser, Dowell, and Clewley (Chap. 5) and He, M. von Davier, Greiff, Steinhauer, and Borysewicz (Chap. 7) discuss problem-solving contexts involving human-agent collaboration. Zhu and Bergner (Chap. 19) discuss the analysis of actions in collaborative games with many players. Learning in the context of an automated tutoring system can be a form of collaboration, and several chapters in this volume address analysis of dialogue data (e.g., Graeser et al., Chap. 5; Griffin, Chap. 8; Olsen, Aleven, & Rummel, Chap. 10) and “joint visual attention” based on dual eye-tracking (Olsen et al., Chap. 10).

1.3 A Taxonomy of Collaborative Problem Solving

In preparing this volume we thought it would be useful to identify some of the key issues associated with collaborative assessment. We did this by proposing a taxonomy of collaborative assessment factors as shown in Fig. 1.1.

The taxonomy identifies four groups of variables to be considered in assessing collaboration. These are participant background variables (cognitive ability, personality, knowledge, demographics, and heterogeneity in backgrounds), task variables (e.g., well- vs. ill-defined tasks, assigned roles, the content domain of the task, whether the task is familiar or novel, and whether the task is a cooperative or competitive one), process variables that can be measured during problem solving, or during learning [e.g., number of statements made, turn taking, personal acknowledgement, goal and planning statements, comprehension monitoring

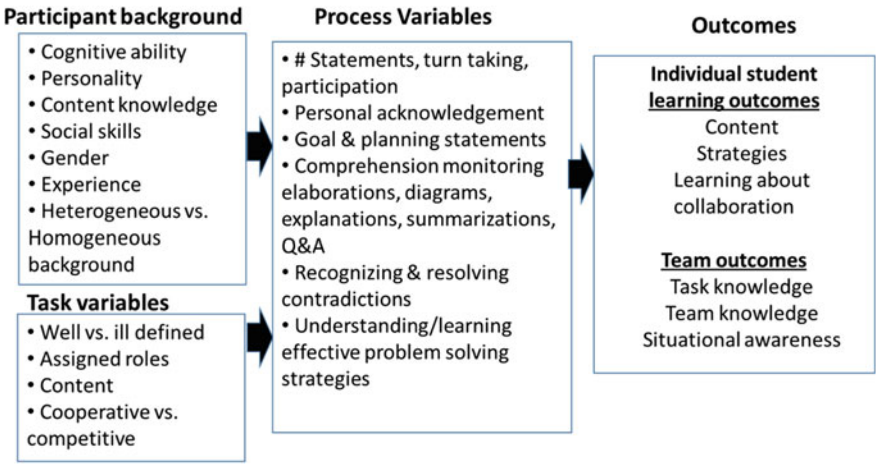

Participant background
• Cognitive ability
• Personality
• Content knowledge
• Social skills
• Gender
• Experience
• Heterogeneous vs. Homogeneous background
Task variables
• Well vs. ill defined
• Assigned roles
• Content
• Cooperative vs. competitive
Process Variables
• # Statements, turn taking, participation
• Personal acknowledgement
• Goal & planning statements
• Comprehension monitoring elaborations, diagrams, explanations, summarizations, Q&A
• Recognizing & resolving contradictions
• Understanding/learning effective problem solving strategies
Outcomes
Individual student learning outcomes
Content
Strategies
Learning about collaboration
Team outcomes
Task knowledge
Team knowledge
Situational awareness

Fig. 1.1 Taxonomy of collaborative assessment factors

(indicated through elaborations, diagrams, explanations, summarizations, questions), recognizing and resolving contradictions, and understanding and learning effective problem-solving strategies]. These variables all would seem to be amenable to coding, and the chapters by Graeser et al. (Chap. 5), Rose et al. (Chap. 6), He et al. (Chap. 7), and Griffin (Chap. 8) suggest various strategies for doing so.

The fourth category in the taxonomy suggests possible individual and collective outcomes. These include knowledge of the topic and problem-solving strategies, and learning about collaboration at the individual level; and task knowledge, team knowledge, and situational awareness at the team level. All of the chapters included or suggested these kinds of outcomes.

We also developed a set of guiding issues we hoped the chapters would address. These included the following:

1. *When is collaboration useful?* Collaboration might not always be the best approach for solving a problem; it might sometimes be better working alone. But a number of chapters identify situations in which collaboration is necessary due to the fact that the problems worked on were large and complex, as they often are in organizations (Salas et al., Chap. 2; Acencio & DeChurch, Chap. 3). In other cases, different participants know different parts of the solution making collaboration necessary (Graeser et al., Chap. 5; He et al., Chap. 7; Griffin, Chap. 8). And collaboration between a teacher and student or between students can be useful in learning (Graeser et al., Chap. 5; Olsen et al., Chap. 10). A related question concerns the specific effects of collaboration on individual knowledge and learning. Is there evidence that students learn new strategies as a result of collaborating? Does working together increase motivation? Several chapters discuss tools to study the effects of collaboration on learning (Graeser et al., Chap. 5; He et al., Chap. 7; Griffin, Chap. 8; Olsen et al., Chap. 10), but it seems little systematic work has addressed this topic thus far.

2. *Why does collaboration sometimes fail?* Collaboration requires a certain degree of coordination between team members, and teams sometimes fail. Failure can be due to interpersonal conflicts, hurt feelings, social loafing and its effects on other members of the team, and also by disagreements about goals, off-topic conversations, and time wasting generally. Several chapters present schemes to code potential behavior along these lines (e.g., Griffin, Chap. 8; Rosé, Howley, Wen, Yang, & Ferschke, Chap. 6; Hao et al., Chap. 9; Bergner, Walker, & Ogan, Chap. 16). Some task environments provide tools to investigate such behavior (e.g., Khan, Chap. 11).
3. *How do we assign individual credit when several individuals are working together?* Credit assignment is one of the thornier problems in collaborative assessment. In PISA 2015 credit assignment was simplified by standardizing the collaboration and having a student work with collaborating agents (Graeser et al., Chap. 5; He et al., Chap. 7). However, Griffin (Chap. 8) and Hao et al. (Chap. 9) suggest that it may be possible to isolate individual contributions even when two humans are working together. Tools widely used in organizational psychology, such as BARS and BOS are designed to identify individual performance.

1.4 The Data

Students or workers in collaborative settings talk, negotiate, hypothesize, revise and respond, orally, with gestures, and on line with chats and emoticons, acronyms, and so on. All of these seem to matter, within the context of collaboration. Data from CPS tasks can be characterized as either (a) individual and team (collective) outcome data, such as the correct/incorrect assessment of an action or task at the individual or team level, or (b) process data. Process data offer insights into the interactional dynamics of team members, which is important both for defining collaborative tasks and for evaluating the results of the collaboration (Morgan, Keshtkar, Graesser, & Shaffer, 2013). Data from collaborative tasks consist of time-stamped sequences of events registered in a *log file*. From a statistical perspective, these activity logs or log files are detailed time series describing the actions and interactions of the users.

A challenge in analyzing log file data is determining the meaning of individual actions and chats. There may be some process variables that are relatively easy to measure, such as participation level of each team member and turn taking. However, beyond these kinds of variables, interpreting actions and chats may be much more challenging due to the dynamics and the sheer volume and complexity of data generated in log files.

Dynamics. In collaborative problem-solving, interactions will change over time and will involve time-lagged interrelationships. If there are two people on a team, the actions of one of them will depend both on the actions of the other and on his or her own past actions. The statistical models needed to accurately describe the dynamics of these interactions bring us outside the realm of traditional psychometric models.

Volume and complexity of data. The challenge of interpreting actions and chats in collaborative interactions in computerized educational environments is that they produce data of extraordinarily high dimensionality (often containing more variables than people for whom those variables are measured). Extracting key features from the noise in such data is crucial to make analysis useful and computationally tractable (Kerr & Chung, 2012).

In addition, with the technological advantages of systems for recording, capturing, and recognition of multimodal data (e.g., Kinect® for Windows, 2016), the data from collaborative interactions contain discourse, actions, gestures, tone, and body language that result in a deluge of data. To these types of data we can add neurophysiological data collected with (portable) EEG headsets (see Stevens et al., Chap. 20). In this volume, several chapters propose different approaches to analyzing the data from teams.

1.5 The Book

The working meeting was organized into panels on *evaluation* (of team performance in organizations; of students working together in tutoring environments, games, and simulations; and of collaborative problem solving in educational settings) and *statistical models* (for dependent process data, and for collaboration and group dynamics). For this edited volume we organize chapters into two major sections: *Part 1: Framework and Methods*, and *Part 2: Modeling and Analysis*.

1.6 Part 1: Framework and Methods

Part 1 includes several chapters that provide overviews, lessons learned, and frameworks for organizing assessment of collaboration and collaborative problem solving from both organizational and educational perspectives. The two perspectives are different. In organizations, teams often are assembled to enable sharing of expertise to optimize organizational performance, and teams are nested in larger teams and organizations in a hierarchical fashion. This leads to a strong emphasis on organizational structure and multilevel coordination. In education, teams are assembled primarily for learning, team members are typically novices rather than experts, and there is relatively more concern for and attention given to dynamics rather than structure and hierarchies. In both organizations and in education there is growing interest in the use of new technologies for collaboration, and in this volume we review several new technologies including dual-eye-tracking, and wearable sensors to measure speech features, body movements, and proximity to others. There also is a particular concern for challenges associated with collaborations in special populations, and here we review those associated with cross cultural collaborations, and collaborations among individuals with disabilities.

1.7 Collaboration from an Organizational Perspective (Chaps. 2–4)

The volume begins with a 30-year perspective on the field of team performance in organizations (Salas et al., Chap. 2). Salas et al. define *teams* as consisting "of two or more people who have defined roles and depend on each other to accomplish a shared goal" (p. 22), and they point out the importance of the multilevel nature of the study of individuals nested in teams nested in organizations. They suggest a taxonomy of methods (self and peer assessments, observations, and objective measures) for assessing both team processes and team outcomes at both individual and team levels. A particularly useful contribution is their Table 2.1 which summarizes key findings and references for team performance measurement over the past 30 years.

Teams themselves can be part of larger multiteam systems in organizations. The multilevel nature of teams is the focus of the next chapter (Acencio & DeChurch, Chap. 3), which defines multiteam systems as two or more teams that work interdependently towards the achievement of collective goals. The importance of this distinction is critical for assessment because individuals contribute to collaboration in modern organizations in different ways. Acencio and DeChurch introduce a vocabulary for relationships in multiteam studies which includes an important but underappreciated distinction between *confluent* (outcomes at one level are consistent with outcomes at another level) and *countervailing* (e.g., positive outcomes at one level, such as team solidarity, are associated with negative outcomes at another level, such as in-group vs. out-group identification) effects. They summarize multiteam system research in their helpful Table 3.1 which gives references for multiteam studies, along with associated predictors, outcomes, and the nature of the relationships examined in those studies.

Teams can certainly be understood at multiple levels, from individuals to groups to organizations. In the next chapter, Fiore (Chap. 4) proposes that advances in technology enhance our ability to address the multiple levels of teamwork and collaboration. As examples he points out that social neuroscience identifies processes such as neural synchrony (compare with Stevens, Steed, Galloway, Lamb, & Lamb, Chap. 20), that neuropeptides are affected by team behavior, sociometric badges and sensor technology enhance understanding of interactions (compare with Khan, Chap. 11), and network analysis and bibliometrics contribute to our understanding of effective collaboration (compare with Zhu, Chap. 19, and Sweet, Chap. 18).

1.8 Collaboration from an Educational Perspective (Chaps. 5–9)

The next four chapters switch to the topic of collaboration in school, rather than in the workplace, and they do so from a variety of perspectives. One form of collaboration is seen in tutorials and dialogues in the context of a learning system, such as computerized instruction. Graeser et al. (Chap. 5) define collaborative competency as "the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills, and efforts to reach that solution" (p. 75). By defining collaboration between agents, assessment developers can control the collaborative context to isolate interactions in a laboratory-like setting. Graeser et al. have done so with a variety of game and simulation tools such as a human-agent tutorial dialogue system called Auto tutor (Graesser, Wiemer-Hastings, Weimer-Hastings, Kreuz, & the Tutoring Research Group, 1999), and a trialogue, featuring conversations between a student and two agents with different roles (e.g., peer, teacher). A useful feature of the chapter is its discussion of the merits of studying collaboration through simulations and dialogues with agents as opposed to between people, for the control it affords.

Analysis of discussions during learning can be studied in a laboratory context as Graeser et al. (Chap. 5) do, but they can also be studied in naturalistic contexts ranging from Massive Open Online Courses (MOOCs) to classrooms, informal learning environments, and Wikipedia collaborations. Rose, Howley, Wen, Yang, and Ferschke (Chap. 6) discuss this approach and propose a set of tools, software resources, and opportunities to participate in the broader community of learning scientists engaged in DANCE (discussion affordances for natural collaborative exchange). Tools include coding schemes to analyze chat and face-to-face discussion data, along cognitive, motivational, and social dimensions; "a publicly available data infrastructure" called DiscourseDB (Ferschke, 2016) designed to facilitate analysis of discussion data across chats, blogs, emails, wikis, and other platforms; and a modeling framework for associating interactions with outcomes.

A collaborative problem solving task based on human-agent collaborations was developed for the PISA 2015 international assessment of 15 year olds, which enables comparisons across over 60 countries. He et al. (Chap. 7) discuss a two-by-two matrix of problem-solving skills (exploring, representing, executing, reflecting) by collaboration skills (understanding, action, organization), which guided assessment development. The assessment itself presented a problem to a student who must work with virtual other students (agents) to solve it. The problem and the chat history remains on the display screen. Preliminary analyses show that the assessment is reliable, and that performance on different problem solving tasks was highly correlated, providing convergent validity evidence. A key message in the chapter is that it is possible to measure collaborative skills in the context of large-scale, multi-language, international assessments.

The collaborative problem solving task in PISA 2015 was a person collaborating with a computer agent, but assessments of person-to-person collaborations are also possible. Two such approaches are discussed in this volume. The first (Griffin, Chap. 8) reviews an approach that grew out of the international Assessment and Teaching of 21st Century Skills (ATC21S) project (Griffin & Care, 2015) in which each problem solver knows something about the problem the other does not, a method known as the jigsaw. Griffin proposes social and cognitive components of collaborative problem solving, with social components of participation (e.g., actions, interactions), perspective taking (responsiveness, adaptation), and social regulation (e.g., negotiation and self-awareness), and cognitive components of task regulation and knowledge building. Griffin and his colleagues have developed a coding system to categorize actions (e.g., chats, keystrokes, mouse movements), and have conducted some preliminary analyses suggesting the validity of the overall approach across diverse language and cultural settings, and the ability to assess the individual contribution to the collective problem-solving effort.

Another person-to-person collaborative framework is proposed in the next chapter (Hao et al., Chap. 9) which allows for disentangling individual from group skills, and cognitive from CPS skills. Hao et al. discuss the challenges of building standardized assessment with CPS tasks and provide strategies to address them. Specifically, they illustrate their recommendations with the Collaborative Science Assessment prototype that includes the Tetralogue (a collaboration between two students and two computer agents) and a collaborative science simulation task. The prototype enables the assessment of science and CPS skills and allows for the collection of fine-grain collaborative process data based on students' chats and actions. The chapter introduces the design of the prototype along with preliminary findings from the first large-sample administration through the crowdsourcing platform, Amazon Mechanical Turk (n.d.) (Mason & Suri, 2012).

1.9 Technology Developments and Collaborative Assessment (Chaps. 10 and 11)

The next two chapters present novel technologies with the potential to expand data collection for studying collaboration. Olsen et al. (Chap. 10) discuss the use of dual eye-tracking, defined as a method in which "eye-tracking data from people working on a task are analyzed jointly" (p. 1XX), particularly when those students are working together while learning with an automated (intelligent) tutoring system (ITS). In particular, they study "joint visual attention," during middle school mathematics lessons on an ITS, which they measure as "the relative amount of time two students are looking at the same area at the same time" (p. 4xx). They show in a series of studies that joint visual attention is higher when there is discussion about the problem-solving, that joint visual attention varies according to the type of collaboration invoked, and that it relates positively to posttest outcomes.

Advanced novel technologies that can be used to discuss collaboration across a range of contexts is the topic of the chapter by Khan (Chap. 11). He discusses recent advances in technologies for multimodal data collection focusing particularly on nonverbal behavior. These include a wide range of audio and visual data capturing technologies, including video cameras and *sociometric badges*, made up of a microphone, accelerometer, Bluetooth position sensors, and infrared line-of-sight sensors, which produce data that can be analyzed to measure high level constructs such as activity (engagement), mimicry (mirroring), and conversational turn taking.

Rapid advances are occurring with these technologies, which are just beginning to be employed in the service of assessing collaboration. Khan's chapter begins to spell out how low-level digital data can be analyzed to assess high-level constructs through a hierarchical framework.

1.10 Collaborative Assessment Issues for Special Populations (Chaps. 12 and 13)

The final two chapters in Part 1 address special issues that have gained increased attention in assessment development generally, and may be particularly important in collaborative assessment. Burke, Feitosa, Salas, and Gelfand (Chap. 12) discuss the importance of cross-cultural contexts, highlighting areas in which cultural diversity could have the largest impact on collaboration. These include cultural variations in perceptions of the power-structure, tolerance of uncertainty, the way cultures make attributions (to individuals or systems), and differences in broad systems of cultural values. Acknowledging cross-cultural differences and identifying ways to mitigate potential conflicts that may interfere with teamwork are essential steps in building fair assessments.

Finally, in the world of testing and assessment there has been a growing interest and attention over the past several years given to issues of accommodations to make testing fair to all test takers (Lovett & Lewandowski, 2015). There are undoubtedly unique challenges in assessing individuals with disabilities in collaboration and collaborative problem solving. This is the topic of the chapter by Hakkinen and White (Chap. 13) who highlight advances in technology, which have enabled advances in universal design, defined as "an approach in which systems are designed at the outset to directly support a broad range of abilities and disabilities." These include assistive technologies, such as screen readers and augmentative communication tools, which now can be blended in smart phones and personal computers. Such advances are now implemented in policy regulations and technology standards for accessibility, particularly regarding Web applications. But the emphasis on technology-based systems in assessing collaboration will require further efforts to avoid exclusion of individuals with disabilities.

1.11 Part 2: Modeling and Analysis

The chapters in the second part of the volume address a wide variety of approaches to modeling collaboration and analyzing collaborative data, including relational events, process data, social network data, and even brain activity. The first four chapters in Part 2 focus on capturing, analyzing, and modeling the interdependences among team members, the next two illustrate the application of social network analysis to team data, and the final chapter demonstrates the value of neurophysiological data in understanding team effectiveness.

1.12 Modeling Interdependencies Among Team Members (Chaps. 14–17)

A characteristic of collaboration is that it is dynamic: One person acts, a second person acts in response to the first person, and then the first person or someone else in turn responds to those actions. This dynamic character of collaborations is challenging from the traditional psychological testing and assessment perspective because standard theory in psychological testing, whether classical test theory or item response theory, is based on the assumption of independence between acts. Under the local independence assumption, two item responses (or test scores) can be correlated, due to a common underlying latent variable, but one response (or one test score) does not affect the next response, other than from the fact that both have a common cause.

Given the centrality of response interdependency in collaboration, several approaches have been proposed to model it directly. One is Relational Events Modeling (REM), introduced by Contractor and Schecter (Chap. 14) for modeling individual interactions over time. REM makes use of the digital traces recorded during the collaborative process, such as a transcript or chat logs, and makes statistical inferences concerning the dynamics in the collaboration, such as one participant's tendency to redo the work of another based on their history of working together. Their chapter provides details on model building and model estimation, and provides an explanatory example. The authors also discuss the potential application of REM as an assessment tool.

A second approach for modeling collaboration views human interaction as coordinated in time, suggesting temporal dependence and temporal clustering for event data, meaning that one's actions affects the probability of another's action. Halpin and A. von Davier (Chap. 15) propose point processes, and the Hawkes process in particular, as a useful statistical framework for modeling temporally clustered data. They illustrate the value of the Hawkes process approach by modeling events in a professional basketball game, specifically, passes (successful vs. unsuccessful) and shots taken (successful vs. unsuccessful) and show how the method detects and models temporal dependence among players' actions.

The chapters by Contractor and Schecter (Chap. 14) and Halpin and A. von Davier (Chap. 15) demonstrate the modeling of dynamic collaborations in work activities and basketball games. In education a common collaboration is tutoring. Peer tutoring is particularly intriguing as a means for exploring collaboration issues because not only is it widely practiced but determining the effectiveness of various collaboration strategies could have widespread benefits. Bergner et al. (Chap. 16) explore peer tutoring collaborations in a "proof of concept" investigation demonstrating the use of dynamic Bayesian network models, specifically, hidden Markov Models (HMM). They explore which of various tutor actions are associated with successful or unsuccessful student results ("model 1") and how student gains are associated with certain patterns of student-tutor activities ("model 2"). The authors discuss the model building and estimation processes, and compare HMMs with logistic regression for predicting outcomes.

The fourth illustration of collaboration modeling in Part 2 is set in the interpersonal context of face-to-face interactions between infants and mothers. Chow, Ou, Cohn, and Messinger (Chap. 17) show how system science methodologies—ones that explicate how system components affect a system's structure and behavior over time—can be used to study interpersonal dynamics. Dyadic processes, such as infant-mother interactions, are characterized by nonstationarities, such as those synchronous, reciprocal influences between infants learning to respond to a mother's emotions which are in turn affected by the infant's actions. This approach can be contrasted with the conventional, and static, stages-of-development view in developmental psychology. The chapter illustrates how sources of nonstationarities can be decomposed and analyzed using spline and nonparametric functions to assist in the understanding of the dynamics during interaction.

1.13 Social Network Analysis (Chaps. 18 and 19)

Social network analysis is a quantitative method for analyzing social structures, such as connections among friends and acquaintances, students in schools, or business colleagues, to study influencing patterns, disease transmission, knowledge sharing, norms, and many other phenomena. Social network analysis thus would seem to be an ideal method for studying collaboration. The next two chapters illustrate advances in social network analysis that expand its usefulness. Sweet (Chap. 18) reviews the application of standard conditionally independent social network models that are currently used in education. She then introduces two hierarchical network models that can apply to more complex situations: The hierarchical latent space model (HLSM) and hierarchical mixed membership stochastic block models (HMMSBM). These are useful for generalizing the findings from a single network, or when the setting of the systems include hierarchy. An advantage of hierarchical network models is that they can capture interdependencies at individual, team, and higher levels.

Zhu and Bergner (Chap. 19) use social networks to model the complex dependencies among teams that share one or more members. They propose the use of bipartite networks in which individuals and teams are represented using different types of nodes and links indicating team membership. They also introduce two analysis tools, the bipartite model in the family of exponential random graph models (ERGMs), and multiple correspondence analysis for bipartite network data. These two methods enable studying team assembly, and the impact of individual and team attributes on performance in collaborative tasks.

1.14 Assessing Team Harmony and Synchrony with Neurophysiological Data (Chap. 20)

All the chapters in Part 2 to this point model behavior, but the final chapter in Part 2 takes one step further to model neurophysiologic data. Stevens et al. (Chap. 20) reviews a study in which they tracked EEG levels of US Navy navigation team members at various stages of a training simulation. In the study they found that high (compared to low) resilience teams showed relatively greater neurodynamic organization during a pre-simulation briefing, but relatively lower neurodynamic organization, indicating more flexibility, during the scenario training segment. They discuss how their approach operationalizes the concepts of *team rhythm*, and being *in-synch*, and demonstrates that physiological data may contribute to an additional level of understanding of how the best teams function as a cohesive unit.

1.15 Conclusions

Building educational assessments entails several requirements:

- a clear definition of the construct,
- a good understanding of the way in which the construct is instantiated in practical demonstrations,
- a careful task development that provides the opportunity for and elicits the appropriate behavior needed to support the claims to be made about someone's skills and abilities,
- a well-designed log file for the fine-grain data from the process and outcomes from complex tasks, and
- appropriate scoring and analyses of these data.

Moreover, educational assessments need to be reliable and valid. Developing collaborative educational assessments is challenging because it is difficult to ensure that all of these requirements are met. Recent advances in technology have begun to allow for major breakthroughs in developing complex and group-worthy collaborative tasks and for collecting time-stamped fine-grain data. Similarly, a fresh way of thinking across and beyond disciplines, from data sciences, to computer science, cognitive psychology, artificial intelligence, and psychometrics has started to open the door to possibilities for accurate predictions of an individual's performance in complex settings. One example of this transdisciplinary work is the introduction of computational psychometrics, a framework in which data-driven methods (machine learning, data mining) and theory-driven methods (psychometrics, statistics, and cognitive sciences) blend, and which allows for and supports the development of next-generation educational assessments, including collaborative assessments (von Davier, 2015, in press; von Davier & Mislevy, 2015). The contributors to this volume reflect upon and discuss the measurement issues of collaboration in their disciplines, often in specific applications.

We hope that the value of this edited volume is in its aiding and encouraging readers to transcend these separate disciplines. Through this edited volume we hope to inspire the search for knowledge across and beyond disciplines to build collaborative educational assessments.

Acknowledgements This book was jointly sponsored by Educational Testing Service and the Army Research Institute for the Behavioral Sciences. The authors thank reviewers James Carlson, Don Powers, and Meghan Brenneman for comments on an earlier version of this article. This work was completed when Alina A. von Davier was employed by Educational Testing Service. The opinions expressed in this paper are those of the authors and not of Educational Testing Service or ACT.

The purpose of this edited volume was to bring together researchers from diverse disciplines—educational psychology, organizational psychology, learning sciences, assessment design, communications, human-computer interaction, computer science, engineering and applied science, psychometrics—who shared a research interest in examining learners and workers engaged in collaborative activity. The collaboration could be as a work team, as a group of students learning together, or as a team working together to solve a problem. There have been several volumes concerned with teamwork and collaboration of workers from an organizational perspective (see Salas, Reyes, & Woods, Chap. 2, Table 2.1) and some research on collaboration in education from a collaborative learning perspective (Care & Griffin, 2014; Griffin & Care, 2015). However, these two broad fields, educational and organizational social science research, have proceeded largely independently despite many shared concerns. Over the last several years some attention has been given to assessment and measurement of 21st century skills, such as teamwork and collaboration, as reflected in several National Research Council reports (2011, 2012, 2015), and special issues of the journals *Applied Measurement in Education* (Greiff & Kyllonen, 2016), and *Journal of Educational Measurement* (A. von Davier, in press).

Given the interest in collaboration and the need to address measurement issues more systematically, Educational Testing Service's (ETS) Research and Development division provided funding for a working meeting, *Innovative Collaborative Assessment*, held in Washington DC, in November 2014.[1] The Army Research Institute joined ETS to support related activities including the preparation of this volume. We organized the working meeting and assembled this volume because of the growing awareness of the importance of collaboration in school and in the workplace coupled with the fact that we do not yet have good methods for assessing it. There clearly is a need for better assessment and better measurement models for collaboration and collaborative skills. It was our shared goal that by assembling this volume we would create synergies among experts from different disciplines, working from different assumptions and perspectives, but able to contribute to an emerging vision on assessing collaboration.

[1] The working meeting is described at http://www.cvent.com/events/innovative-assessment-of-collaboration-two-day-working-meeting/custom-19-4110888121994d93bccb78007a50ebc8.aspx.

References

Amazon Mechanical Turk Requester Tour. (n.d.). Retrieved from https://requester.mturk.com/tour

Bartram, D. (2013). Scalar equivalence of OPQ32: Big five profiles of 31 countries. *Journal of Cross-Cultural Psychology, 44,* 61–83.

Burrus, J., Elliott, D., Brenneman, M., Markle, R., Carney, L., Moore, G. … Roberts, R. D. (2013). *Putting and keeping students on track: Toward a comprehensive model of college persistence and attainment* (Research Report 13–14). Princeton, NJ: Educational Testing Service.

Care, E., & Griffin, P. (2014). An approach to assessment of collaborative problem solving. *Research and Practice in Technology Enhanced Learning, 9*(3), 367–388.

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. ERIC Number: ED519465, ISBN-0-8237-0888-8. Washington, DC: Partnership for 21st Century Skills. Retrieved from http://www.p21.org/storage/documents/FINAL_REPORT_PDF09-29-06.pdf

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136,* 1092–1122.

Deming, D. J. (2015). The growing importance of social skills in the labor market (Working Paper 21473). *National Bureau of Economic Research.* Retrieved from http://www.nber.org/papers/w21473

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support Army selection and classification decisions (Technical Report 1311).* Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Duhigg, C. (2016, February 28). What Google learned in trying to build the perfect team. *New York Times Magazine*, MM20.

Ferschke, O. (2016). *DiscourseDB core wiki*. https://github.com/DiscourseDB/discoursedb-core.wiki.git

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the Tutoring Research Group. (1999). Auto tutor: A simulation of a human tutor. *Journal of Cognitive Systems Research, 1,* 35–51.

Greiff, S., & Kyllonen, P. C. (in press). Contemporary assessment challenges: The measurement of 21st century skills (Guest Editors' Introduction). *Applied Measurement in Education, 29*(4), 243–244.

Griffin, P., & Care, E. (Eds.). (2015). *Assessment and teaching of 21st century skills: Methods and approach*. Dordrecht, the Netherlands: Springer.

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100). New York, NY: Guilford Press.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big-five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). New York, NY: Guilford Press.

Kerr, D., & Chung, G. K. W. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining, 4*(1), 144–182.

Kinect® for Windows. (2016). *Meet Kinect for Windows*. https://developer.microsoft.com/en-us/windows/kinect. Microsoft.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*(1), 191–207.

Korn Ferry International. (2014–2016). *Leadership architect technical manual* (Item number 82277). Minneapolis, MN: Author. http://static.kornferry.com/media/sidebar_downloads/KFLA_Technical_Manual.pdf

Kyllonen, P. C., & Bertling, J. P. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277–285). Boca Raton, FL: CRC Press.

Lovett, B. J., & Lewandowski, L. J. (2015). *Testing accommodations for students with disabilities: Research-based practices*. Washington, DC: American Psychological Association.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's mechanical turk. *Behavioral Research, 44*(1), 1–23.

Morgan, B., Keshtkar, F., Graesser, A., & Shaffer, D. W. (2013). Automating the mentor in a serious game: A discourse analysis using finite state machines. In C. Stephanidis (Ed.), *Proceedings of the 15th international conference on human-computer interaction (HCI international)* (pp. 591–595). Berlin, Germany: Springer.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640–647.

National Association of Colleges and Employers. (2014). *The skills/qualities employers want in new college graduate hires.* Retrieved from http://www.naceweb.org/about-us/press/class-2015-skills-qualities-employers-want.aspx

National Center for Education Statistics. (2014, September 29). *NAEP innovations symposium: Collaborative problem solving*. Washington, DC: Author.

National Research Council. (2011). *Assessing 21st century skills: Summary of a workshop* (J. A. Koenig, Rapporteur). Committee on the Assessment of 21st Century Skills. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century* (J. W. Pellegrino & M. L. Hilton, Eds.). Committee on Defining Deeper Learning and 21st Century Skills. Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2015). *Enhancing the effectiveness of team science* (N. J. Cooke & M. L. Hilton, Eds.). Committee on the Science of Team Science. Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Oh, I.-S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology, 96*(4), 762–773.

Organisation for Economic Cooperation and Development. (2013). *PISA 2015: Draft collaborative problem solving framework*. Paris, France: Author. Retrieved from https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81*(3), 524–539.

Salgado, J. F., & Tauriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*(1), 3–30. doi:10.1080/1359432X.2012.716198

von Davier, A. A. (2015, July). *Virtual and collaborative assessments: Examples, implications, and challenges for educational measurement*. Invited Talk at the Workshop on Machine Learning for Education, International Conference of Machine Learning, Lille, France http://dsp.rice.edu/ML4Ed_ICML2015

von Davier, A. A. (in press). Computational psychometrics in support of collaborative assessments. In A. A. von Davier (Ed.). Measurement issues in collaborative learning and assessment [Special Issue]. *Journal of Educational Measurement*.

von Davier, A. A., & Mislevy, R. J. (in press). Design and modeling frameworks for 21st century: Simulations and game-based assessments. In M. Falkner-Bond & C. Wells (Eds.), *Educational measurement: From foundations to future*. New York, NY: Guilford.

Wang, L., MacCann, C., Zhuang, X., Liu, O. L., & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students. *Canadian Journal of School Psychology, 24*(2), 108–124.

Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement, and subgroup differences. *Human Performance, 17,* 433–461. doi:10.1207/s15327043hup1704_5.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19,* 188–202.

Weinberger, C. J. (2014). The increasing complementarity between cognitive and social skills. *The Review of Economics and Statistics, 96*(5), 849–861. doi:10.1162/REST_a_00449. http://www.mitpressjournals.org/doi/abs/10.1162/REST_a_00449

Zu, J., & Kyllonen, P. C. (2012, April). Item response models for multiple-choice situational judgment tests. In *Situational Judgment Testing for Educational Applications*. Symposium conducted at the meeting of the National Council of Measurement in Education, Vancouver, Canada.

Part I
Framework and Methods

Chapter 2
The Assessment of Team Performance: Observations and Needs

Eduardo Salas, Denise L. Reyes, and Amanda L. Woods

Abstract The abundance of teams within organizations illustrates the importance of team performance measurement—tools that measure teamwork. Taking into account the inherently complex nature of teams, this chapter presents a few insights and a picture of the research and practice on teamwork measurement over time. We define what makes a team and identify the characteristics of an effective team. Then, we present critical observations to team performance measurement that reflect the 30 years of experience of the first author, at observing, measuring, and assessing team performance in various domains. These observations provide insight into what attitudes, behaviors, and cognitions—how teams feel, act, and think—play an integral role in performance assessment, while taking situational factors and construct considerations into account. Support is presented from the literature on teams and performance measurement, and we provide major contributions from a sample of team performance measurement literature in the past 30 years. We conclude with a discussion on needs for developing future team-based measurement approaches. In this discussion of the future, emphasis is placed on our need, as a field, to continue closing the gap between research and practice through designing and validating effective performance-based measures that target practitioner needs.

Keywords Teamwork · Performance measurement · Assessment

E. Salas (✉) · D.L. Reyes · A.L. Woods
Rice University, Houston, USA
e-mail: Eduardo.Salas@rice.edu

D.L. Reyes
e-mail: Denise.L.Reyes@rice.edu

A.L. Woods
e-mail: Amanda.L.Woods@rice.edu

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_2

2.1 Introduction

Teams are a way of life in organizations. The military, the aviation and space industry, healthcare, corporations, and educational institutions all depend and rely on teams today more than ever. Effective teamwork creates knowledge, minimizes errors, promotes innovation, saves lives, enhances productivity, increases job satisfaction, and ensures success. Teams, when deployed, trained, and led correctly, can be powerful. But insuring that teams perform, learn, develop, and mature is not easy. In fact, it is complex and difficult. A key component to help with this is performance measurement—tools that measure teamwork. Thus we need to create these tools to accurately determine the strengths and weaknesses of the team. This is not an easy goal. We need valid, reliable, theory-driven practices that account for the dynamic nature of teams (Brannick & Prince, 1997; Langan-Fox, Wirth, Code, Langfield-Smith, & Wirth, 2001). This is a tall order, but progress has been made—much progress; this volume is a testament of that progress.

This chapter contributes to the volume by presenting a few insights and a picture of the research and practice on measuring teamwork over time. We will first provide some definitions to set the stage. We will next present some critical observations about measuring team performance. These observations are based primarily on the 30 years of experience of the first author at observing, measuring, and assessing team performance in various domains. We also rely on the literature to support these observations. Lastly, we will discuss some needs for developing future team-based measurement approaches.

2.2 Some Definitions

A team consists of two or more people who have defined roles and depend on each other to accomplish a shared goal (Salas, Dickinson, Converse, & Tannenbaum, 1992). In order to understand how teams work and subsequently perform, we have to understand how much the team knows, what skills they possess, and the overall attitude that they bring to the table; we refer to these elements as *team competencies* (Rosen et al., 2008).

The nature of teams is inherently complex, because individual workers are nested in teams, which are nested in organizations (Cannon-Bowers & Salas, 1997; Cannon-Bowers, Tannenbaum, Salas, & Volpe, 1995). With teams adding this dynamic layer of complexity, it is critical to slice apart and analyze what characteristics are embedded in the team, as well as the various factors (e.g., individual, team, and organizational factors) that contribute to team performance (Marks, Mathieu, & Zaccaro, 2001). The first step to understand team performance is to identify what characteristics the team possesses starting out. Examples of these *inputs* are individual motivation, attitudes, and personality traits (Driskell, Salas, & Hughes, 2010). Team-level inputs include power distribution, cohesion, and team resources (Marks et al., 2001). However, inputs are not limited to these characteristics. The type of task and how complicated it is also play a role. Next, we have to identify the *processes*, or the actions that occur when the team is working together to complete a task (LePine, Piccolo, Jackson, Mathieu, & Saul, 2008; Marks et al., 2001). Thus, it is apparent that teams are riddled with complexity, even at their nascent stages.

Though assessing team performance is challenging, we do it because team performance is linked to team effectiveness. Salas, Stagl, Burke, & Goodwin, (2007) defined team effectiveness as the result of a judgment process whereby an output is compared to a subjective or objective standard. Essentially, the results of the team's inputs and processes are evaluated. Therefore, to ensure accuracy, we must match the outcome with the correct methods of measurement (Rosen, Wildman, Salas, & Rayne, 2012). The team yields outcomes at the team and individual levels. Team-level outcomes require the effort of all team members, such as coordination and communication. Individual-level outcomes include a team member's attitude toward the team, which is related to team performance. Organizational-level outcomes are the resulting products of the task and how the team impacts the overall organization. Before we move on, it is important to remember that individual changes in attitude, motivation, mental models, and task knowledge, skills, and attitudes (KSAs) can impact future team processes and performance outcomes, because individuals make up a team (Cannon-Bowers et al., 1995; Tannenbaum, Beard, & Salas, 1992). Taking all these factors into consideration, in order for us to improve performance assessment, we must adopt a multilevel approach (individual, team, and organizational) to understand all the elements contributing to the way team members work together and what they produce based on their actions. With all of these issues in mind, we will now present our observations (in no particular order).

2.3 Observations

2.3.1 Observation 1: We Know a Lot

Team performance measurement is not a perfect science, yet. However, we have learned a great deal over the past 30 years, and we have amassed a robust body of literature on this area of measurement in an effort to address issues that researchers and practitioners face (Brannick & Prince, 1997; Cooke, Kiekel, & Helm, 2001; Kozlowski & Bell, 2003; Rosen et al., 2012; Wildman et al., 2012). Rosen and colleagues (2013) elucidated key components of team performance, as well as providing helpful guidelines for assessment in the context of performance in healthcare settings. Kendall and Salas (2004) addressed methodological concerns by investigating reliability and validity issues impacting team performance metrics. Taking a finer lens to team processes, He, von Davier, Greiff, Steinhauer, and Borysewicz (2015) have made significant progress towards the development of assessments (e.g., the Programme for International Student Assessment [PISA]) that capitalize on current technology to capture team collaborative problem-solving abilities. Due to recent research efforts, the ability to objectively capture real-time performance is also on the horizon (Stevens, Galloway, Lamb, Steed, & Lamb, 2017). To summarize, we know about why, how, when, and what to measure, but gaps remain. We will talk more on this later; for a more in-depth glimpse into team performance measurement advances, refer to Table 2.1.

Table 2.1 Sample of team performance measurement literature in the past 30 years

Source	Major contribution(s)
Kendall and Salas (2004)	Examined the criterion problem of team performance, explained current methods for measuring teamwork, and addressed issues of reliability and validity of the measures
Cooke et al. (2000)	Reviewed methods for measuring team knowledge (cognition), a component of teamwork skills and provided methodological needs for the measurement
Salas, Priest, and Burke (2005)	Discussed perceived challenges for those who are responsible for the development of team performance measurement systems, which include defining the purpose, selecting suitable scenarios to use, accounting for timing of the measurement, quantifying responses of teams, and determining how to simplify the collection of data
Salas, Burke, Fowlkes, and Priest (2004)	Explained research in a style for organizations to understand more about the basic elements of team performance measurements, practical requirements for evaluating teamwork skills, and tools to evaluate team skills in order to implement measurements in applied settings
Fowlkes et al. (1994)	Developed Targeted Acceptable Responses to Generated Events or Tasks (TARGETS), an event-based approach for measuring behaviors in teams
Rosen et al. (2013)	Defined key elements of team performance. Provided a guide for measuring, assessing, and diagnosing team performance in healthcare systems
Rosen et al. (2012)	Addressed the challenges faced in measuring team dynamics in real-world settings. Defined team performance measurement methods are defined and presented best practices for developing practical measurements
Salas, Burke, and Fowlkes (2005)	Provided a brief overview of team performance measurement over the past 20 years. Provided a taxonomy of teams present in organizations along with the challenges of measuring their performance. Discussed how these challenges are currently being addressed and offered practical suggestions for practitioners

2.3.2 Observation 2: Context and Purpose of Measurement Matter

There is no "silver bullet" when creating a team performance measurement tool. We need to think about the context when creating all aspects of a measurement; who, how, and what is being used to conduct the evaluation. Team size, complexity of the task, physical environment of the task, task interdependence, and the amount of communication and interaction required to complete the task should also be considered (Salas, Burke, & Fowlkes, 2005).

The purpose of the performance measurement (i.e., team feedback) should determine what will be collected, and what needs to be collected should determine what kinds of resources are being used for the measurement (Meister, 1985). When choosing a team performance measurement, it is important to remember that all measures need adjustments and modifications in order to have a suitable quality for the required purpose (Salas et al., 2015). Targeting the idiosyncrasies within the team will give you a better idea of what modifications need to be made.

2.3.3 Observation 3: It Is Best to Triangulate

When it comes to measuring teamwork, it is nearly impossible to collect all of the necessary data from just one source. As noted by Dickinson and McIntyre (1997), "it surely takes a group or team of observers to obtain the necessary information to measure all instances of teamwork" (p. 37). There are a number of ways in which data can be collected. One can use self-report, peer assessments, observations, and objective outcomes. Using different types of data collection is optimal for getting the most data. It is best to use a combination of both qualitative and quantitative data. Subjective ratings are subject to bias; however, there are ways to reduce this bias. For example, observer ratings need to involve interrater reliability to make sure that the variable is being rated accurately from the beginning to the end (Rosen et al., 2012). We can do this by randomly selecting sessions for more than one rater to code and then comparing their ratings (Shrout & Fleiss, 1979). Also, different raters can focus on different areas based on their expertise. For example, supervisors can be used for summative assessments, while peers or subordinates can rate for ongoing or developmental evaluations.

Since teamwork is performed by individuals, it is also important to measure team performance at the individual level. We can achieve a more accurate evaluation of team performance when it is measured at multiple levels. Analysis at the individual level can pinpoint the members who effectively demonstrate teamwork skills (e.g., leadership, coordination, communication). Also, measuring both processes and outcomes can extend the amount of information you can learn about the team's performance. Looking at processes can give you diagnostic information that addresses issues of development and can serve as a guide for feedback.

Outcome measures, on the other hand, can provide you with "bottom line" performance. Making sure that you have a triangulation approach to collecting data can help ensure validity and address the limitations of the approaches when they are used alone. You do not want any potentially useful data to go unnoticed!

2.3.4 *Observation 4: Team Size Matters*

Teams come in all shapes and sizes. When it comes to performance, the size of the team can actually make a difference (Dyer, 1984; Sundstrom, De Meuse, & Futrell, 1990). Hackman (1987) suggested having teams with the least amount of people that are necessary to perform the task. The more team members that are added to a group, the lower the cohesion (McGrath, 1984) as well as group performance (Nieva, Fleishman, & Reick, 1978). The size of a team can be determined by the task at hand or the type of team (i.e., human-computer, distributed teams).

Larger teams run into issues of less flexibility and more differences within the team. More people mean more individual differences. These challenges also carry into the way the team's performance is measured. Team performance measurement for large teams should include contingency planning, implicit coordination during task execution (i.e., shared mental models), information management, developed understanding of subteams, and an assessment of intra- and interteam cooperation. When conducting observations in complex team settings, raters should not observe more than two team members. This helps to avoid overlooking interactions (Dickinson & McIntyre, 1997).

2.3.5 *Observation 5: Subject Matter Experts Can Assess Only Four or Five Constructs*

Experts cannot assess or distinguish more than five team-based constructs. Measuring a construct requires subject matter experts (SMEs), who are individuals that have a strong understanding of the task setting and must make judgments about different team-based constructs. There is a tendency for observers and practitioners alike, to measure all they can measure—sometimes 12 to 14 constructs! Again, raters cannot distinguish these constructs; they all correlate at the end. Our experience is that raters should be trained to focus on only four or five constructs to avoid redundancy (Smith-Jentsch, Zeisig, Acton, & McPherson, 1998). When more than five related constructs are examined, the dimensions start to overlap and become more correlated with each other, making practical distinctions among teams almost impossible. In this case, less is better. Therefore it is wise to select team-based constructs carefully and use only those that matter for team performance.

2.3.6 Observation 6: It Is Best to Capture the ABCs—Attitudes, Behaviors, and Cognitions

It is best to capture representative attitudes, behaviors, and cognitions of teamwork. Teamwork has all of these elements. Noting Observation 5 above, it is best to choose one or two relevant ABCs to capture. Fortunately, an extensive body of research exists surrounding essential ABCs that promote effectiveness. This provides a clear outline of what measurement should capture. Recently, team orientation has been identified as a core attitudinal component of high performing teams (Salas, Sims, & Burke, 2005). Effective teams also promote a wide variety of behaviors such as communication, coordination, and cooperation, to name a few (Campion, Medsker, & Higgs, 1993; Kozlowski & Bell, 2003). For a more in-depth look at team behaviors, refer to Rousseau, Aubé, and Savoie (2006).

Regarding team cognition, shared mental models play an important role in ensuring that team members are on the same page. Successful development of shared mental models helps aggregate the knowledge of each member on the team to create a common understanding of what, how, and when the team needs to accomplish a goal or task (Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000). For further discussion of team cognition and its component parts, refer to DeChurch and Mesmer-Magnus (2010). Taken as whole, capturing ABCs is critical for determining how to best measure a team and maximize performance outcomes.

Capturing attitude is commonly used for measuring team performance because it is easy and does not rely on many resources. Measuring attitude is as simple as having team members individually answer a set of items, using a Likert scale to express their feelings in regard to particular statements. Recently, we have also seen examples of attaining information signals by capturing facial expressions, gestures, posture, and periods of silence (Anders, Heinzel, Ethofer, & Haynes, 2011; Shippers, Roebroeck, Renken, Nanetti, & Keysers, 2010; Schokley, Santana, & Fowler, 2003; Stevens et al., 2017). We need to measure attitudes because they are associated with team performance (Hackman, 1990; Peterson, Mitchell, Thompson, & Burr, 2000). In regard to behaviors, these can easily be captured through observation. We will elaborate more on what behaviors need to be observed in Observation 7. As for team cognition (knowledge), it still remains a challenge to find a promising method to measure this construct, but it is important to measure because it affects performance (Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015). In a methodological review, Cooke, Salas, Cannon-Bowers, and Stout (2000) explained that we need to go beyond typical assessments to understand the structure of team knowledge. Different aspects of measurement for this construct include elicitation method (e.g., self-report, eye tracking, communication analysis), team metric, and aggregation method. Nonetheless there is a lot more to be done in regard to measuring cognition (e.g., Wildman et al., 2012).

2.3.7 Observation 7: Behavioral Markers Matter

Behavioral markers are paramount in performance measurement (Flin & Martin, 2001). Accurately capturing observable behaviors within a team is critical to assessing a team's attributes. These markers should be studied in the context of the environment in which they are being applied. However, mapping constructs to the environment is only a part of the battle. Behavioral markers must be specific and the constructs of interest need to be clearly defined. We have already touched upon various widely used measurement tools that rely on observable team behavior in Observation 6, but a more granular lens must be used to establish what behaviors are of interest. To accurately execute this, time should be taken to methodically carry out the subsequent steps. First, we must establish the behaviors of interest. Next, we must systematically map constructs onto the behaviors. Additionally, we must clearly define the identified constructs. Finally, we must contextualize the behavioral markers by assessing them in the actual performance environment.

2.3.8 Observation 8: It's All About the Constructs, Not the Method!

A primary issue surrounding constructs is the heightened emphasis placed on the method at the expense of unique traits present in the team. It is important to remember that all teams are not equal! Teams possess both explicit (e.g., observable behaviors such as verbal communication) and implicit qualities (e.g., unobservable processes such as shared mental models; Entin & Serfaty, 1999; Rosen et al., 2012). Due to the developmental nature of teams, certain phenomena (e.g., implicit qualities) emerge in teamwork that can be difficult to capture. Research has attempted to overcome this challenge by placing primary emphasis on the tools used to assess teamwork, but this can sacrifice important aspects of teamwork that influence performance. Most available tools are limited to assessing observable behaviors, but some of the team's most important interactions are implicit and therefore difficult to capture. To illustrate this, in an operating room a patient goes into cardiac arrest; a nurse immediately hands the surgeon necessary tools while the anesthesiologist monitors the patient's current condition and the surgeon attempts to stabilize the patient. This is a good example of a scenario in which implicit coordination is key to the success of the surgical team. Many of these actions need to take place in a matter of seconds; the actions are highly interdependent and do not require explicit communication. As you can imagine, it would be difficult to measure how aligned the team's shared mental model was or how this impacted their ability to coordinate in a highly stressful situation.

Another challenge that centers on the constructs involved in measurement is the statistical method used in analysis. Though accurate and appropriate statistical analysis is critical to team assessment, it does not sufficiently capture performance all by itself. Many methods of analysis exist that establish the reliability and validity of constructs, but researchers should proceed with caution so as not to become completely reliant on these analyses. The environment and situation being assessed should also play a critical role, to ensure that empirical constructs translate to practical settings (Rosen et al., 2012).

Taking these factors into account when defining constructs is crucial to developing accurate and adaptable performance measures specific to the team. When it comes to teams, adaptability is key (Rosen et al., 2013) and should be reflected in the measurement process. Contextualization should, again, be taken into account, aligning constructs with team competencies to provide accurate construct definitions (Cannon-Bowers et al., 1995).

2.3.9 Observation 9: Measurement of Teamwork Is Not a "One-Stop Shop" Dynamic Phenomenon

Adding to the complexity of teamwork is the simultaneous need for multiple measurement methods that address the episodic nature of team processes. Teams do not run on fixed intervals; they accomplish different tasks at different times. Hence, it is important to recognize that there is no universal form of measurement that captures performance (Rosen et al., 2012), but keen observation can be a powerful tool when selecting a form of assessment (Rosen et al., 2012).

Although it might be a labor-intensive process to obtain these data, there are new unobtrusive approaches that are promising for team performance measurements. The most popular approaches for observing behavior are event-based measurement, real-time assessment, classification schemes, coding, and behavioral rating scales.

Event-based measurement plays out a scenario where the training objectives are connected to what exactly needs to be assessed. This lets the assessor design events specific to the behaviors to be evaluated. Having control over the events enhances the measurement reliability. Two measurement tools that were developed using the event-based approach are targeted acceptable responses to generated events or tasks (TARGETS; Fowlkes, Lane, Salas, Franz, & Oser, 1994) and team dimensional training (TDT; Smith-Jentsch et al., 1998). One of the most common approaches uses behavioral rating scales such as the behaviorally anchored rating scales (BARS), introduced by Smith and Kendall (1963). Other rating scales include behavioral observation scales (BOS) and graphic rating scales (Latham & Wexley, 1977; Patterson, 1922).

For capturing performance, assessment tools should take on a multilevel perspective (e.g., individual, team, and organizational levels), to accommodate the changes that teams encounter through their life cycle (Rosen et al., 2012; Wildman et al., 2012). Performance should also be measured frequently through a variety of techniques to prevent method bias. However, a challenge this poses is the overuse of dimensions or measures. Frequently assessing a team can get in the way of team dynamics or otherwise alter the team's normative behavior.

Unobtrusive measures are useful in situations where a team's performance is constantly changing, because they do not disrupt the workflow of the team members. The electroencephalography (EEG) approach to capturing team performance has also shown promise in regard to being unobtrusive while allowing for the real-time assessment of behaviors (Stevens et al., 2017). Automated performance measures (e.g., sociometric badges and audio recording devices) have also shown promise with regard to being both unbiased and unobtrusive. Expounding further on the area of automation, PISA made strides towards capturing both cognitive and social aspects of collaborative problem solving through computer-based assessment (He et al., 2015). One caveat about this method is that automated performance measures are not "stand alone" measures. They still need to be coupled with nonautomated forms of measurement. However, this need for multiple measures holds true for many forms of performance assessment.

2.3.10 Observation 10: What Is Good for Science Is Not Necessarily Good for Practice

Bridging the gap between research and practice is a critical focus for assessing teamwork performance. This is a challenge because what is good for team research is not always what practitioners want. Researchers can assess many elements of teamwork performance in a controlled laboratory setting, but this freedom can cause researchers to lose sight of what is relevant for practice. Practitioners need tools that are unobtrusive, diagnostic, economical, and easy to use (Rosen et al., 2012). Researchers do not always take an approach that meets these needs. This disparity between research and practice is compounded by the inconsistencies that exist within the dimensions of theoretical teamwork models.

2.3.11 Observation 11: Don't Ignore the Basics

It is important to go back to the basics to ensure good practice. The underlying premise behind successful measurement provides a sound foundation for future research and practice efforts. The basics illustrate the guiding principles, emerging trends, and considerations for team performance measurement. Outlining clear constructs that target the attitudes, behaviors, and cognitions pertinent to teamwork, while factoring in the context of the environment, lays the groundwork for effective performance measurement.

Great strides have been made in the area of teamwork performance measurement. An area that shows great promise in particular has been modeling and simulation (Fiore, Cuevas, Scielzo, & Salas, 2002; Hao, Liu, von Davier, & Kyllonen, 2015). However, more development is needed to maintain focus as we move forward. Some of the challenges that still remain are determining what to measure, developing reliable instruments that are diagnostic, and ensuring that these instruments can be implemented across the life span of the team, while placing a heavy emphasis on practicality. To ensure that new methods of assessment are grounded in a reliable and valid foundation, we must go "back to the basics."

2.4 The Future

2.4.1 Observation 12: We Need Tools that Capture the ABCs of Teamwork Dynamically in Real Time that Are Pragmatic, Relevant, and Unobtrusive

This is the holy grail of team measurement. That is the next step in the future. Efforts have been made to reach this goal; see promising work in Table 2.2. Future research should aim at improving the effectiveness of team measurement, such as the work being done by Cooke (2015), who noted that measuring interactions can

Table 2.2 Overview of observations on team performance measurements

Team performance measure observations		References
Context matters	– No perfect protocol, technique, or format exists – All need adjustments and modifications – All teams are not created equal	Meister (1985) Salas, Priest, and Burke (2005)
Best to triangulate	– Use self-report, peer assessments, and observations – It takes a team to evaluate a team – Use multiple angles, facets, and components	Dickinson and McIntyre (1997) Rosen et al. (2012) Shrout and Fleiss (1979)
Team size matters	– Size of the team makes a difference – Team size affects performance and how performance measures are implemented	Dyer (1984) Sundstrom et al. (1990) Hackman (1987)
SMEs can only assess four or five constructs	– The more constructs, the more correlated they are – Rater training helps – Observations help in debriefing	Smith-Jentsch et al. (1998)

(continued)

Table 2.2 (continued)

Team performance measure observations		References
Best to capture ABCs	– New unobtrusive approaches are promising – Low-level metrics are also promising – Cognitions remain a challenge	Smith-Jentsch et al. (1998) Peterson et al. (2000) Liu et al. (2015)
Behavioral markers matter	– Be specific – Define constructs of interest precisely – Take time and be systematic – Contextualize constructs	Flin and Martin (2001) Kendall and Salas (2004)
Need to focus on constructs	– Discipline to define constructs is lacking – Lots of focus on the statistics technique is necessary but not sufficient – Obsession with methodological tool often comes at the expense of the phenomena	Cannon-Bowers et al. (1995) Rosen et al. (2012)
Measuring teamwork is a dynamic phenomenon	– Teams do different things at different times – Measure often – Unobtrusive measures are needed	Rosen et al. (2012) Wildman et al. (2012)
What is good for science is not necessarily good for practice	– Practitioners need simple, easy to use, relevant, and diagnostic measures – Researchers can sometimes afford to throw in the "kitchen sink"	Rosen et al. (2012)
Don't ignore the basics	– Guiding principles are often ignored – New emerging approaches are needed – More is needed, so we should go back to basics	Salas, Priest, and Burke (2005) Morgan, Glickman, Woodard, Blaiwes, and Salas (1986)
We need tools that capture the ABCs of teamwork dynamically in real-time that are pragmatic, relevant, and unobtrusive	– Aim at improving the effectiveness of team measurement – More unobtrusive measures are needed – Acknowledge the advancement of technology and increased usage of online tools for assessment	Awwal et al. (2015) Cooke (2015)

easily be done unobtrusively and that more unobtrusive measures are needed. Research also needs to acknowledge the advancement of technology and increased usage of online tools for assessment (Awwal, Griffin, & Scalise, 2015).

2.5 Conclusion

It is evident that team performance measures are important throughout many industries, and since not all teams are created equally, it is important to modify the measurement based on the specific team. When a measurement system is developed, it should address the question: *Why do we measure?* This question requires a clear definition of the purpose of the measurement tool (von Davier & Halpin, 2013). The purpose behind measuring performance is to generate research, provide teams with feedback, develop team training, evaluate performance, and plan for the future.

During the development of measurement tools another question you need to answer is: *What areas of performance should be captured?* As previously described, to accurately assess performance, the team should be measured on multiple dimensions and the conceptual elements of the measure should be clearly defined. This leads into the temporal considerations of performance assessment: *When should we measure?*Teamwork should be assessed midway through the performance cycle as well as after the conclusion of the performance episode. This begs the question: *Where should teamwork performance be measured?* Teamwork should be measured both in the field through the use of unobtrusive measures as well as in a synthetic environment (Rosen et al., 2013). Lastly, the proper method of analysis should be selected: *How should we measure performance?* Teamwork should be captured through self-report measures, observation, simulations, and balanced scorecards (Rosen et al., 2013).

Acknowledgements This work was supported in part by contract NNX16AB08G with the National Aeronautics and Space Administration (NASA) and contract NBPF03402 with the National Space Biomedical Research Institute (NSBRI) with Rice University. The views expressed in this work are those of the authors and do not necessarily reflect the organizations with which they are affiliated or their sponsoring institutions or agencies.

References

Anders, S., Heinzle, J., Weiskopf, N., Ethofer, T., & Haynes, J. (2011). Flow of affective information between communicating brains. *Neuroimage, 54,* 439–446.

Awwal, N., Griffin, P., & Scalise, S. (2015). Platforms for delivery of collaborative tasks. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: methods and approach* (pp. 105–113). Dordrecht, Netherlands: Springer.

Brannick, M. T., & Prince, C. W. (1997). An overview of team performance measurement. In M. T. Brannick, E. Salas, & C. W. Prince (Eds.), *Team performance assessment and measurement: Theory, methods, and applications* (pp. 3–18). Mahwah, NJ, USA: Lawrence Erlbaum.

Campion, M. A., Medsker, G. J., & Higgs, A. C. (1993). Relations between work group characteristics and effectiveness: Implications for designing effective work groups. *Personnel Psychology, 46,* 823–847.

Cannon-Bowers, J. A., & Salas, E. (1997). A framework for developing team performance measures in training. In M. T. Brannick, E. Salas, & C. W. Prince (Eds.), *Team performance assessment and measurement: Theory, methods, and applications* (pp. 45–62). Mahwah, NJ, USA: Lawrence Erlbaum.

Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E., & Volpe, C. E. (1995). Defining competencies and establishing team training requirements. In R. A. Guzzo & E. Salas (Eds.), *Team effectiveness and decision making in organizations* (pp. 333–380). New York, NY, USA: Wiley.

Cooke, N. J. (2015). Team cognition as interaction. *Current directions in psychological science, 24*(6), 415–419.

Cooke, N. J., Kiekel, P. A., & Helm, E. E. (2001). Measuring team knowledge during skill acquisition of a complex task. *International Journal of Cognitive Ergonomics, 5*(3), 297–315.

Cooke, N. J., Salas, E., Cannon-Bowers, J. A., & Stout, R. J. (2000). Measuring team knowledge. *Human Factors, 42*(1), 151–173.

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: A meta-analysis. *Journal of Applied Psychology, 95*(1), 32–53.

Dickinson, T. L., & McIntyre, R. M. (1997). A conceptual framework for teamwork measurement. In M. T. Brannick, E. Salas, & C. W. Prince (Eds.), *Team performance assessment and measurement: Theory, methods, and applications* (pp. 19–43). Mahwah, NJ, USA: Lawrence Erlbaum.

Driskell, J. E., Salas, E., & Hughes, S. (2010). Collective orientation and team performance: development of an individual differences measure. *Human Factors, 52*(2), 316–328.

Dyer, J. L. (1984). Team research and team training: A state of the art review. In F. A. Muckler (Ed.), *Human factors review* (pp. 285–323). Santa Monica, CA, USA: Human Factors Society.

Entin, E. E., & Serfaty, D. (1999). Adaptive team coordination. *Human Factors, 41*(2), 312–325.

Fiore, S. M., Cuevas, H. M., Scielzo, S., & Salas, E. (2002). Training individuals for distributed teams: Problem solving assessment for distributed mission research. *Computers in Human Behavior, 18*(6), 729–744.

Flin, R., & Martin, L. (2001). Behavioral markers for crew resource management: A review of current practice. *The International Journal of Aviation Psychology, 11*(1), 95–118.

Fowlkes, J. E., Lane, N. E., Salas, E., Franz, T., & Oser, R. (1994). Improving the measurement of team performance: The TARGETs methodology. *Military Psychology, 6*(1), 47–61.

Hackman, J. R. (1987). The design of work teams. In J. Lorsch (Ed.), *Handbook of organizational behavior* (pp. 315–342). New York, NY, USA: Prentice Hall.

Hackman, J. R. (1990). *Groups that work (and those that don't)*. San Francisco, CA: Jossey-Bass.

Hao, J., Liu, L., von Davier, A. A., & Kyllonen, P. (2015). Assessing collaborative problem solving with simulation based tasks. *International Society of the Learning Sciences Proceedings.*

He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., & Borysewicz, P. B. (2015). *Collaborative problem solving measures in the programme for international student assessment (PISA).*

Kendall, D. L., & Salas, E. (2004). Measuring team performance: Review of current methods and consideration of future needs. In J. W. Ness, V. Tepe, & D. R. Ritzer (Eds.), *The science and simulation of human performance* (Vol. 5, pp. 307–326). Bingley, UK: Emerald Group Publishing.

Kozlowski, S. W. J., & Bell, B. S. (2003). Work groups and teams in organizations. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 333–375). London, UK: Wiley.

Langan-Fox, J., Wirth, A., Code, S., Langfield-Smith, K., & Wirth, A. (2001). Analyzing shared and team mental models. *International Journal of Industrial Ergonomics, 28*(2), 99–112.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30,* 255–268.

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. (2008). A meta-analysis of teamwork processes: Tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology, 61*(2), 273–307.

Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, D. (2015). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen & S. Ferrara (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 344–359). Hershey, PA, USA: Information Science Reference.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review, 26,* 356–376.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology, 85*(2), 273–283.

McGrath, J. E. (1984). *Groups: Interaction and performance*. Upper Saddle River, NJ, USA: Prentice Hall.

Meister, D. (1985). *Behavioral analysis and measurement methods*. New York, NY, USA: Wiley.

Morgan, B. B., Jr., Glickman, A. S., Woodard, E. A., Blaiwes, A. S., & Salas E. (1986). *Measurement of team behaviors in a Navy environment* (Technical Report No. NTSC TR-86-014). Orlando, FL, USA: Naval Training Systems Center.

Nieva, V. F., Fleishman, E. A., & Reick, A. (1978). *Team dimensions*: *Their identity, their measurement, and relationships* (Final Tech. Rep., Contract DAH19-78-C-0001). Washington, DC, USA: Advanced Resources Research Center.

Patterson, D. G. (1922). The Scott Company graphic rating scale. *Journal of Personnel Research, 1*, 361–376.

Peterson, E., Mitchell, T. R., Thompson, L., & Burr, R. (2000). Collective efficacy and aspects of shared mental models as predictors of performance over time in work groups. *Group Processes & Intergroup Relations, 3*(3), 296–316.

Rosen, M. A., Salas, E., Wilson, K. A., King, H. B., Salisbury, M., Augenstein, J. S. … Birnbach, D. J. (2008). Measuring team performance in simulation-based training: Adopting best practices for healthcare. Simulation in Healthcare, 3(1), 33–41.

Rosen, M. A., Schiebel, N., Salas, E., Wu, T., Silvestri, S., & King, H. (2013). How can team performance be measured, assessed, and diagnosed. In E. Salas & K. Frush (Eds.), *Improving patient safety through teamwork and team training* (pp. 59–79). New York, NY, USA: Oxford University Press.

Rosen, M. A., Wildman, J. L., Salas, E., & Rayne, S. (2012). Measuring team dynamics in the wild. In A. Hollingshead & M. S. Poole (Eds.), *Research methods for studying groups: A guide to approaches, tools, and technologies* (pp. 386–417). New York, NY, USA: Taylor & Francis.

Rousseau, V., Aube, C., & Savoie, A. (2006). Teamwork behaviors: A review and an integration of frameworks. *Small Group Research, 37*(5), 540–570.

Salas, E., Benishek, L., Coultas, C., Dietz, A., Grossman, R., Lazzara, E., et al. (2015). *Team training essentials: A research-based guide*. New York, NY, USA: Routledge.

Salas, E., Burke, C. S., & Fowlkes, J. E. (2005). Measuring team performance "in the wild:" Challenges and tips. In W. Bennet Jr., C. E. Lance, & D. J. Woehr (Eds.), *Performance measurement*: *Current perspectives and future challenges* (pp. 245–272). Mahwah, NJ, USA: Erlbaum.

Salas, E., Burke, C. S., Fowlkes, J. E., & Priest, H. A. (2004). On measuring teamwork skills. In J. C. Thomas & M. Hersen (Eds.), *Comprehensive handbook of psychological assessment* (Vol. 4, pp. 427–442). Hoboken, NJ, USA: Wiley.

Salas, E., Dickinson, T., Converse, S. A., & Tannenbaum, S. I. (1992). Toward an understanding of team performance and training. In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 3–29). Norwood, NJ, USA: Ablex.

Salas, E., Priest, H. A., & Burke, C. S. (2005b). Teamwork and team performance measurement. In J. R. Wilson & N. Corlett (Eds.), *Evaluation of human work* (3rd ed., pp. 793–808). Boca Raton, FL, USA: CRC Press.

Salas, E., Sims, D. E., & Burke, C. S. (2005c). Is there a “big five” in teamwork? *Small Group Research, 36*(5), 555–599.

Salas, E., Stagl, K. C., Burke, C. S., & Goodwin, G. F. (2007). Fostering team effectiveness in organizations: Toward an integrative theoretical framework. *Nebraska Symposium on Motivation, 52,* 185–243.

Schippers, M., Roebroeck, A., Renken, R., Nanetti, L., & Keysers, C. (2010). Mapping the information flows from one brain to another during gestural communication. *Proceedings of the National Academy of Sciences USA, 107,* 9388–9393.

Shockley, K., Santana, M.-V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance, 29*(2), 326–332.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Smith-Jentsch, K. A., Zeisig, R. L., Acton, B., & McPherson, J. A. (1998). Team dimensional training: A strategy for guided team self-correction. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 271–297). Washington, DC, USA: American Psychological Association.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*(2), 149–155.

Stevens, R., Galloway, T., Lamb, J., Steed, R., & Lamb, C. (2017). Linking team neurodynamic organizations with observational ratings of team performance. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration*. New York, NY, USA: Springer.

Sundstrom, E., De Meuse, K. P., & Futrell, D. (1990). Work teams: Applications and effectiveness. *American Psychologist, 45*(2), 120–133.

Tannenbaum, S. I., Beard, R. L., & Salas, E. (1992). Team building and its influence on team effectiveness: An examination of conceptual and empirical developments. *Advances in Psychology, 82,* 117–153.

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations (ETS Research Report No RR-13-41).* Princeton, NJ, USA: Educational Testing Service.

Wildman, J. L., Thayer, A. L., Pavlas, D., Salas, E., Stewart, J. E., & Howse, W. R. (2012). Team knowledge research: Emerging trends and critical needs. *Human Factors, 54*(1), 84–111.

Chapter 3
Assessing Collaboration Within and Between Teams: A Multiteam Systems Perspective

Raquel Asencio and Leslie A. DeChurch

Abstract Developing assessment methods that capture an individual's capability to collaborate can look to the team and multiteam systems literature, which identifies six critical components of collaboration. These six include team affect/motivation, team interaction processes, and team cognition, as well as corresponding constructs at the system level, multiteam affect/motivation, between-team interaction, and multiteam cognition. This chapter defines and distinguishes teams and multiteam systems and discusses the importance of that distinction for assessing individual collaborative capacity in both small stand-alone teams and larger systems of teams working toward superordinate goals. Particularly, we describe confluent and countervailing forces—the notion that what enables team functioning and effectiveness may or may not also enable the multiteam system effectiveness. Assessments of individual contributions to team and multiteam dynamics must consider the implications to functioning both within and between teams.

Keywords Teams · Multiteam systems · Individual assessment · Confluent and countervailing forces

Teams are now one of the most basic units through which we accomplish tasks, and this reality has important implications for assessment. As many of the most pressing and complex problems are the province of specialized individuals working in teams, assessment methods are needed that enable the measurement of knowledge, skills, abilities, and other experiences (KSAOs) that enable an individual to effectively contribute to team effectiveness. Furthermore, there is mounting evidence that as knowledge becomes increasingly specialized, teams must rely on

R. Asencio (✉)
Purdue University, 403 W State St., West Lafayette, IN 47901, USA
e-mail: rasencio@purdue.edu

L.A. DeChurch
Northwestern University, Evanston, USA

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_3

other teams in order to bring together a greater array of expertise. These larger collectives are called multiteam systems (MTSs), and consist of two or more teams. A defining feature of an MTS is that each team pursues its own proximal team goals, while also working as a larger system of teams, who are interdependent with regard to a more distal superordinate goal (DeChurch & Zaccaro, 2010). Thus, MTSs work in an environment that necessitates attention to both team and MTS functioning. However, although research in this area is still growing, the literature on MTSs is not currently considering the effectiveness of the teams and the MTS at the same time (DeChurch & Zaccaro, 2013). This chapter considers the multiteam structure and its implications for individual assessment.

The context of MTSs brings to light an important duality between the team and the system. On the one hand, individuals in the MTS must focus on team effectiveness. The structure of an MTS is such that teams pursue their own proximal goals. This requires interactions that promote team effectiveness (McGrath, 1984). In addition to a focus on team interactions, individuals must also manage intergroup relations that build the foundation for MTS-level interaction processes to develop, and thus aid with MTS effectiveness. Therefore, individuals are embedded in two groups (i.e., a team and an overarching multiteam system) that require their focused attention and efforts.

Ideally, what enables the effectiveness of the team would also enable the effectiveness of the MTS. However, this may not be the case. The processes that lead to effective teams may not be aligned with the processes that lead to an effective MTS. The notion of confluent and countervailing forces captures both of these situations in MTSs. *Confluent forces* are those in which processes and properties have the same consequence at the team and MTS level of analysis. *Countervailing forces* are those processes and properties that have divergent consequences at the team and MTS levels of analysis (DeChurch & Zaccaro, 2013).

Researchers must strive to incorporate a complete view of MTSs and consider the impact of team- and MTS-level properties on outcomes at multiple levels of analysis. The question then becomes the following: Which processes and properties are important to team and MTS functioning? Furthermore, how do individual KSAOs combine to impact these processes and properties?

In the current chapter we (a) define MTSs and describe the unique characteristics of these teamwork structures, (b) describe aspects of teamwork critical for team and MTS functioning, and (c) describe the notion of confluent and countervailing forces and the implications for individual assessment of collaboration within and between teams.

3.1 Defining Multiteam Systems

For years, organizations have seen the value of assembling teams to leverage the distinct expertise of individual members, who together can achieve optimal solutions. The study of teams and team dynamics has flourished in the fields of industrial organizational psychology and organizational behavior (DeChurch & Mesmer-Magnus, 2010; DeChurch, Mesmer-Magnus, & Doty, 2013; de Wit, Greer, & Jehn, 2012; Gully, Incalcaterra, Joshi, & Beaubien, 2002; LePine, Piccolo, Jackson, Mathieu, & Saul, 2008; Mesmer-Magnus & DeChurch, 2009; Mullen & Cooper, 1994; Stajkovic, Lee, & Nyberg, 2009). However, the increase in globalization has changed the landscape of organizational work. Global work has created a need for teams to reach across organizational and geographic boundaries to work with other teams to solve important environmental, social, technological, and medical issues.

In the same way that individual expertise is brought to bear on a problem within a single team, these complex problems often require the effort of multiple teams that together have the requisite expertise necessary to tackle important issues (DeChurch & Zaccaro, 2010). Collectives composed of tightly coupled teams are called MTSs. MTSs are formally defined as the following:

> Two or more teams that interface directly and interdependently in response to environmental contingencies toward the accomplishment of collective goals. MTS boundaries are defined by virtue of the fact that all teams within the system, while pursuing different proximal goals, share at least one common distal goal; and in doing so exhibit input, process, and outcome interdependence with at least one other team in the system. (Mathieu, Marks, & Zaccaro, 2001, p. 290)

There are five important key features of MTSs that are implied in the definition put forth by Mathieu et al. (2001). First, MTSs are composed of a minimum of two teams. These component teams are “non-reducible and distinguishable wholes” (p. 291), that have proximal goals and interdependent members. Second, in addition to proximal goals, component teams share a common superordinate goal for which all teams are collectively responsible. Third, the structure or configuration of the MTS is determined by the goals, performance requirements, and technologies adopted. The performance environment determines what goals need to be accomplished by both the component teams and the MTS. The goals for the system are organized into a hierarchy with proximal team goals at the lowest level and distal MTS goals at the highest level (Mathieu et al., 2001; Zaccaro, Marks, & DeChurch, 2012). Fourth, MTSs are larger than teams, but smaller than the embedding organization(s). While MTSs can be housed within the same organization (known as *internal MTSs*), an MTS may cross formal organizational boundaries (known as *cross-boundary MTSs*). The fifth key feature of MTSs is that component teams have input, process, or outcome interdependence with at least one other team in the MTS. The type of interdependence in an MTS is intensive, with component teams working in a reciprocal manner, or closely with one another (Zaccaro et al., 2012). By contrast, pooled interdependence, in which teams work in isolation and “pool” their outputs, or sequential interdependence, in which teams work in succession of one another, are not typically characteristic of a tightly coupled system of teams.

3.2 Boundary Issues in Multiteam Systems

Mathieu et al. (2001) conceptualized MTSs as entities that are larger than teams, but smaller than organizations. One contention about the MTS structure is that they could be simply considered as large teams, with at best, subunits that characterize different groups (DeChurch & Mathieu, 2009). However, in MTSs, component teams are loosely coupled so that, although tied to other teams through interdependence, the team boundary remains intact. Indeed, it is valuable to consider the reciprocal influence of component teams and the MTS, much in the same way that we consider the impact of individuals on a team, and vice versa (Chen & Kanfer, 2006; DeShon, Kozlowski, Schmidt, Milner, & Wiechmann, 2004). This suggests that component teams have their own *entitativity* (Campbell, 1958). The degree of entitativity is the extent to which a group can be considered to be a stand-alone entity. Campbell discussed three factors that determine the entitativity of a group: *proximity*, *similarity*, and *common fate*.

The principle of proximity states that elements that are close together are likely considered to be part of the same group (Campbell, 1958). A component team may be colocated in the same organization, establishing proximity among the members. An example of component teams with high proximity are those in an emergency response MTS. Each component team in the system (e.g., police, fire fighter, emergency medical technician) is colocated within its own brick-and-mortar organization. However, globalization has made virtual teams more prevalent and thus, component teams may also be spread across geographical boundaries. For example, in a large scientific MTS, a component team may be composed of members from different research institutions. Therefore, proximity may only be sufficient to establish entitativity for collocated teams.

The principle of similarity states that elements with similar qualities and characteristics are likely to considered part of the same group (Campbell, 1958). In an MTS this could translate into component teams having specialized roles or functions. For example, in a product development MTS, component teams carry out various functions, such as project management, research and design, programming, data analytics, and marketing. Within each team there are different priorities, languages, and frames of reference, helping to establish each team as a separate unit. However, similarity may not be sufficient to establish entitativity, as component teams in an MTS may serve very similar or overlapping functions. For example, DeChurch and Mathieu (2009) described a firefighting MTS composed of teams with various functions (e.g., fire suppression, ventilation, and search and rescue). In a multialarm fire, there may be several teams with the same function active at the same time (e.g., two search-and-rescue teams).

The principle of common fate states that elements with common processes and outcomes are likely to be considered as part of the same group (Campbell, 1958). Observing the covariation of activities within and across groups, we consider that entitativity is established when the covariation is greater within, rather than across teams (DeChurch & Zaccaro, 2013). In MTSs the goal hierarchy can establish common fate among members of the same team. Although component teams in the MTS share common fate through the accomplishment of an overarching goal, each team has its own team-level goals and priorities (Mathieu et al., 2001). Thus, the commonality of activities and goals within a team is greater than the commonality across teams. Common fate, therefore, is a defining characteristic that serves to differentiate the teams in a system (DeChurch & Zaccaro, 2013).

Assuming that entitativity is established for each component team in the MTS, members must deal with the draw of two foci: the team and the system. Each will have pull on an individual's attention and direct efforts. Managing the team and MTS boundary requires a focus on team and MTS effectiveness, as well as team- and MTS-level goals, making the MTS a complex environment within which members must interact and function. Thus, individual assessments aimed at uncovering an individual's capacity for teamwork must account for these two levels of collaboration.

3.3 Tripartite Taxonomy of Team and MTS Functioning

To clearly establish an understanding of an individual's capacity for collaboration in the context of MTSs, assessments should explore how individual KSAOs contribute to critical facets of teamwork at both the team and MTS levels. While there are many models and taxonomies of teamwork, there is substantial convergence on the notion of three core mechanisms of teamwork: *affect/motivation*, *behavior*, and *cognition* (Kozlowski & Ilgen, 2006; Salas, Rosen, Burke, & Goodwin, 2009).

Team affect/motivation captures aspects of the team or MTS that stem from members' emotions, attachment, and/or motivation. Team cohesion, the result of all of the forces acting upon the individual to remain in the group (Cartwright, 1968; Festinger, Schacter, & Back, 1950), is perhaps the quintessential aspect of team affect. Other affective/motivational constructs include team potency, collective efficacy, and team goal commitment (Gully et al., 2002; Stajkovic et al., 2009). Whereas most studies of team affect/motivation have focused on affect or motivation within relatively small teams, these constructs are meaningful at the larger MTS level as well. Recent dissertations have explored cohesion (DiRosa, 2013) and efficacy (Jimenez-Rodriguez, 2012) at the MTS level.

Team behavior reflects "what teams do" (Kozlowski & Ilgen, 2006, p. 95). Team processes are the verbal and behavioral mechanisms through which individuals combine their effort to accomplish a team task (Cohen & Bailey, 1997). A validated (LePine et al., 2008) taxonomy of team process behaviors was advanced by Marks et al. (2001). This taxonomy details 10 interaction processes needed by individuals as they pursue collective goals. Three preparatory processes include setting goals, analyzing the task, and setting up plans and contingency plans. Four action processes include monitoring progress, monitoring and backing up teammates, monitoring the performance environment, and coordination. Whereas the first two sets of processes meet the task needs of the group, a third set of interpersonal processes allow the group to manage the social context of the team. These interpersonal processes include motivating and confidence building, conflict management, and affect/emotion management.

Whereas the taxonomy was developed for application to small teams, it has been extended to MTSs, and it provides a useful framework for understanding the between-team processes needed when teams share goals with other teams and must collaborate externally. Two initial studies of MTSs adapted several of these processes, such as coordination (DeChurch & Marks, 2006), to the between-team level. Each of the 10 (Marks et al., 2001) processes can be defined at the intra- and inter-team levels, both of which are useful criteria on which to validate individual assessment metrics.

Team cognition captures a team's (or MTS's) organized knowledge (Klimoski & Mohammed, 1994). Interest in the notion of team cognition began in earnest in the 1980s, and progressed in two relatively orthogonal lines of thought. The first observed that individuals who work together develop differentiated systems of encoding and retrieving information. Termed *team transactive memory systems* (TMS; Liang, Moreland, & Argote, 1995; Moreland, 1999; Moreland, Argote, & Krishnan, 1996), this form of team cognition involves two components. First, team members distribute who knows what information so that the team can increase its collective working memory capacity. The second component of TMS is a shared awareness of who knows what. This latter aspect of the construct enables team members to be efficient in their retrieval and allocation of information within the team. The TMS construct has been shown into be a strong mechanism of team effectiveness (Austin, 2003; DeChurch & Marks, 2006; Lewis, Lange, & Gillis, 2005; Littlepage et al. 2008).

The second line of inquiry on team cognition is the concept of a *team mental model*. Team mental models were discovered while observing that expert teams were able to seamlessly coordinate their actions, anticipating one another's needs without the need for communication (Cannon-Bowers, Salas, & Converse, 1993). Subsequent team mental model research has examined a variety of content domains and forms (Klimoski & Mohammed, 1994). Two popular content domains are task work models and teamwork models. The former details the critical aspects of the task and their interrelation; the latter details aspects of needed member interaction and social functioning. Regardless of the content domain, this research generally distinguishes between the similarity and accuracy of team mental models. Interestingly, research finds both similarity and accuracy contribute uniquely to team performance (Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000). Thus, even a shared but inaccurate mental model provides some benefit to team performance.

In sum, decades of team effectiveness research have revealed how teams can be most effective. The general belief is that teams need strong affective/motivational and cognitive states, and behavioral processes in order to function at an optimal level (Kozlowski & Ilgen, 2006; McGrath, 1984). Indeed, meta-analyses confirm the importance of information sharing (Mesmer-Magnus & DeChurch, 2009), cognition (DeChurch & Mesmer-Magnus, 2010), cohesion (Beal, Cohen, Burke, and McLendon, 2003), team processes (LePine et al., 2008), and conflict (De Dreu & Weingart, 2003; DeChurch et al., 2013) for performance as well as other aspects of team functioning. Research has also indicated the importance of some of these aspects of teamwork for MTS effectiveness, thereby revealing those aspects of teamwork (i.e., affect/motivation, behavior, and cognition) that are important for the success of the team and the MTS, respectively.

These extensions are important when addressing teams in the context of MTSs. However, still missing are individual assessments that predict which aspects of the individual contribute to the functioning of both the team and the MTS. Thus, there are two level of complexity to address. First, there is the possibility that factors that contribute to individual performance may be different from those that contribute performance in a team (von Davier & Halpin, 2013). Individual assessment therefore, must account for the team context when determining what individual-level factors contribute to the success of a team. Second, the factors that contribute to team performance may be different from those that contribute to MTS performance (DeChurch & Zaccaro, 2013). It is not enough to simply consider how individual KSAOs impact team functioning or MTS functioning, respectively. Indeed, to get a more complete picture of collaboration within and between teams, assessments must consider the impact that individual KSAOs have on team and MTS functioning simultaneously.

3.4 Confluent and Countervailing Forces

Countervailing forces occur when a process or property manifested at one level (i.e., team or MTS) has opposing consequences at different levels of analysis (DeChurch & Zaccaro, 2013). For example, teams that engage all members in the planning and strategizing phases of a task may encourage participation, empowerment, and buy-in (Lanaj, Hollenbeck, Ilgen, Barnes, & Harmon, 2013), but this type of decentralization across teams may result in coordination failures when there are too many members engaged in cross-team planning. Thus, a team process may have a positive (or negative) effect on an outcome at the team level, and the opposite effect with an outcome at the system level (DeChurch & Zaccaro, 2013). This point is critical for assessment, because validating metrics on one level or the other is deficient in capturing the ways that individuals contribute to collaboration in modern organizations. However, while MTS researchers have acknowledged the potential for countervailing forces in MTSs, virtually none have empirically examined these relationships. Instead, MTS researchers have mainly focused on assessing the homology of team-level relationships at the MTS level, uncovering processes and properties that are helpful or harmful to MTS effectiveness (ignoring team effectiveness). Thus, most hypotheses tested in extant research on MTSs give only part of the story.

Countervailing forces are different from confluent forces, in which a process or property manifested at one level has the same effect on outcomes at both the team and MTS levels. For example, planning activities across teams helps the system to establish a strategy for achieving MTS level goals, but when team and MTS goals are closely aligned, planning between teams can also aid individual teams in developing a strategy for moving forward with team goals. When team and MTS processes are confluent, assessment efforts can validate assessment methods against the consequences at either level.

Table 3.1 summarizes the empirical studies of MTSs. The table lists the various studies conducted on MTSs and the relationships examined. We categorize the nature of the relationships reported in the research. Single-level studies include predictors and criteria at the MTS level of analysis. Multilevel homology studies include predictors and criteria at least two levels of analysis, but hypothesize and test only single-level relations at each level with the aim of discovering the degree to which these relations (e.g., the relation between coordination and performance) are the same at multiple levels of analysis. Confluent and countervailing relationships are specific types of cross-level relations (DeChurch & Zaccaro, 2013). These cross-level relations are relevant to assessment efforts because they can reveal cases where a process has opposite consequences at two levels of analysis. In particular, where an individual characteristic may contribute to a process or property that benefits the team (or MTS), it may do so at the cost of MTS (or team) progress.

As an illustration, imagine that team cohesion exhibits a countervailing effect. Decades of primary studies on a wide variety of teams have shown a strong positive link between cohesion and performance (Beal et al., 2003). While the direction of the relationship has been widely debated, we can generally conclude that teams whose members are emotionally connected to the team tend to be the high performing teams (and vice versa). DiRosa (2013) posited and tested the idea that team cohesion, while generally good for team outcomes, may have detrimental effects at the system level. When teams become insular, it can activate social categorization processes and suppress information sharing and collaboration across teams, effectively undermining MTS performance. Hence, it is important that measures that assess collaboration consider individual contributions to both team- and system-level functioning. Using the example of team cohesion, an individual that contributes to very strong team cohesion may inadvertently set up the perfect conditions for intense intergroup competition. In terms of collaborative capability, such individuals may ultimately lead the team to victory while simultaneously leading the MTS to defeat.

Table 3.1 shows under a dozen empirical studies of MTSs (published at the time of writing this chapter). Whereas most examine both team and multiteam processes as predictors of MTS effectiveness, four of these did not include team-level predictors—meaning they cannot account for the incremental validity of MTS functioning beyond that predicted at the team level. Also relevant to assessment, only one of these studies (Davison, 2012) predicted criteria at both the team and MTS level. Such dual-level studies are needed to properly inform assessment research that will ultimately need to validate predictors on these multilevel criteria. Meta-analytic accumulation across studies can partially compensate for these blind spots in the primary literature.

Table 3.1 Relationships examined in previous MTS research

Study	Predictor level	Predictors examined	Criterion level	Criteria examined	Relationships examined			
					Single level	Multilevel homology	Confluent	Countervailing
Davison, Hollenbeck, Barnes, Sleesman, & Ilgen (2012)	Team; MTS	Coordinated action	MTS	Performance	X			
Davison (2012)	Team; MTS	Roles; goal commitment; identity	Team; MTS	Performance	X	X		X
DeChurch and Marks (2006)	Team; MTS	Strategy and coordination training; coordination; team performance	MTS	Functional leadership; coordination; performance	X			
de Vries, Walter, Van der Vegt, &, Essens (2014)	MTS	Coordination	Team	Performance	X			
DiRosa (2013)	Team; MTS	Interdependence; boundary spanning; goal alignment; cohesion	MTS	Cohesion; goal alignment; readiness	X			
Firth, Hollenbeck, Miles, Ilgen, & Barnes (2015)	Team; MTS	Frame-of-reference training; coordination	MTS	Performance	X			
Jimenez-Rodriguez (2012)	MTS	Efficacy; information sharing uniqueness and openness; transactive memory; trust; shared mental model; communication retrievability; media richness	MTS	Information sharing uniqueness and openness; performance; transactive memory; shared mental models	X			

(continued)

Table 3.1 (continued)

Study	Predictor level	Predictors examined	Criterion level	Criteria examined	Relationships examined			
					Single level	Multilevel homology	Confluent	Countervailing
Lanaj et al. (2013)	MTS	Decentralized planning; planned and actual proactivity, aspirations, and risk-seeking; coordination failures	MTS	Planned and actual proactivity, aspirations, and risk-seeking; coordination failures; performance	X			
Marks, DeChurch, Mathieu, Panzer, & Alonso (2005)	Team; MTS	Action and transition process; interdependence	MTS	Performance; action process	X		X	
Mathieu, Maynard, Taylor, Gilson, & Ruddy (2007)	Team; MTS	MTS coordination; openness climate; team interdependence, team processes	Team	Team process; performance	X		X	
Murase, Carter, DeChurch, & Marks (2014)	MTS	Interaction mental model accuracy; coordination; strategic communication	MTS	Coordination; performance	X			

However, traditional ways of thinking about team and MTS effectiveness preclude us from considering the inherent complexity of teamwork in MTSs, and how individual characteristics play a role in shaping the process of teamwork. The confluence and countervailance perspective provides a more complete understanding of the forces at play that impact both team and multiteam outcomes, both of which need to be considered in developing useful assessment methods.

3.5 Implications

As research on MTSs continues to grow, it is important that researchers begin to take on a more complex view of MTSs. To better enable the success of both team and MTS goals, research should use the confluence/countervailance lens to understand what factors facilitate and impede team and MTS effectiveness. The role of individual assessment in this cause is twofold. First, individual assessment needs to determine what individual-level factors shape an individual's collaborative capacity. For example, researchers may explore the collaborative interactions and the features of successful collaboration (von Davier & Halpin, 2013). Second, individual assessment needs to determine how individuals not only contribute to and shape team-level interactions, but also how individual-level factors may in also influence interactions in the MTS. Further, it is important to consider team and MTS outcomes simultaneously.

3.6 Conclusion

For many teams, MTSs represent a context that imposes new challenges in teamwork. The growing body of literature on MTSs has examined factors that may improve or hinder MTS performance. However, as MTSs are composed of entitative teams with their own local goals, it stands to reason that research should explore the factors that may mutually impact both team and system effectiveness. The present manuscript lays out this framework as a way to validate measures of individual collaborative capability.

Acknowledgements We thank our research collaborators, especially Noshir Contractor and Steve Zaccaro, who have generously contributed to many of the ideas in this chapter through insightful discussions. The preparation of this chapter was supported by the Army Research Institute for the Social and Behavioral Sciences under contract W5J9CQ12C0017, and the Army Research Office under contract W911NF-14-1-0686. The views, opinions, and/or findings contained in this report are those of the authors, and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

References

Austin, J. (2003). Transactive memory in organizational groups: The effects of content, consensus, specialization, and accuracy in group performance. *Journal of Applied Psychology, 88*(5), 866–878.

Beal, D. J., Cohen, R. R., Burke, M. J., & McLendon, C. L. (2003). Cohesion and performance in groups: A meta-analytic clarification of construct relations. *The Journal of Applied Psychology, 88*(6), 989–1004.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science, 3*(1), 14–25.

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In N. J. Castellan (Ed.), *Individual and group decision making: Current issues* (pp. 221–246). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cartwright, D. (1968). The nature of group cohesiveness. *Group Dynamics: Research and Theory, 91,* 109.

Chen, G., & Kanfer, R. (2006). Toward a systems theory of motivated behavior in work teams. *Research in Organizational Behavior, 27,* 223–267.

Cohen, S. G., & Bailey, D. E. (1997). What makes teams work: Group effectiveness research from the shop floor to the executive suite. *Journal of Management, 23*(3), 239–290.

Davison, R. B. (2012). *Implications of regulatory mode and fit for goal commitment, identity and performance in the domain of multiteam systems*. Dissertation, Michigan State University.

Davison, R. B., Hollenbeck, J. R., Barnes, C. M., Sleesman, D. J., & Ilgen, D. R. (2012). Coordinated action in multiteam systems. *Journal of Applied Psychology, 97*(4), 808–824.

De Dreu, C. K., & Weingart, L. R. (2003). Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. *Journal of Applied Psychology, 88*(4), 741.

de Vries, T. A., Walter, F., Van der Vegt, G. S., & Essens, P. J. (2014). Antecedents of individuals' interteam coordination: Broad functional experiences as a mixed blessing. *Academy of Management Journal, 57*(5), 1334–1359.

de Wit, F. R., Greer, L. L., & Jehn, K. A. (2012). The paradox of intragroup conflict: A meta-analysis. *Journal of Applied Psychology, 97*(2), 360–390.

DeChurch, L. A., & Marks, M. A. (2006). Leadership in multiteam systems. *Journal of Applied Psychology, 91*(2), 311–329.

DeChurch, L. A., & Mathieu, J. E. (2009). Thinking in terms of multiteam systems. In E. Salas, G. F. Goodwin, & C. S. Burke (Eds.), *Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches* (pp. 267–292). New York: Taylor & Francis).

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: A meta-analysis. *Journal of Applied Psychology, 95,* 32–53.

DeChurch, L. A., Mesmer-Magnus, J. R., & Doty, D. (2013). Moving beyond relationship and task conflict: Toward a process-state perspective. *Journal of Applied Psychology, 98*(4), 559–578.

DeChurch, L. A., & Zaccaro, S. J. (2010). Perspectives: Teams won't solve this problem. *Human Factors, 52*(2), 329–334.

DeChurch, L. A., & Zaccaro, S. J. (2013). *Innovation in scientific multiteam systems: Confluent & countervailing forces*. Paper presented to the National Academy of Sciences Committee on Team Science. Washington DC: National Academy of Sciences. Retrieved April 5, 2016, from http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_083773.pdf

DeShon, R. P., Kozlowski, S. W., Schmidt, A. M., Milner, K. R., & Wiechmann, D. (2004). A multiple-goal, multilevel model of feedback effects on the regulation of individual and team performance. *Journal of Applied Psychology, 89*(6), 1035–1056.

DiRosa, G. (2013). *Emergent phenomena in multiteam systems: An examination of between-team cohesion.* Dissertation, George Mason University.

Festinger, L., Schacter, S., & Back, K. W. (1950). *Social pressures in informal groups*. Stanford, CA: Stanford University Press.

Firth, B. M., Hollenbeck, J. R., Miles, J. E., Ilgen, D. R., & Barnes, C. M. (2015). Same page, different books: Extending representational gaps theory to enhance performance in multiteam systems. *Academy of Management Journal, 58*(3), 813–835.

Gully, S. M., Incalcaterra, K. A., Joshi, A., & Beaubien, J. M. (2002). A meta-analysis of team-efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. *Journal of Applied Psychology, 87*(5), 819–832.

Jimenez-Rodriguez, M. (2012). *Two pathways to performance: Affective and motivationally driven development in virtual multiteam systems*. Dissertation, University of Central Florida.

Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management, 20*(2), 403–437.

Kozlowski, S. W., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest, 7*(3), 77–124.

Lanaj, K., Hollenbeck, J., Ilgen, D., Barnes, C., & Harmon, S. (2013). The double-edged sword of decentralized planning in multiteam systems. *Academy of Management Journal, 56*(3), 735–757.

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. (2008). A meta-analysis of teamwork processes: tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology, 61*(2), 273–307.

Lewis, K., Lange, D., & Gillis, L. (2005). Transactive memory systems, learning, and learning transfer. *Organization Science, 16*(6), 581–598.

Liang, D. W., Moreland, R., & Argote, L. (1995). Group versus individual training and group performance: The mediating role of transactive memory. *Personality and Social Psychology Bulletin, 21*(4), 384–393.

Littlepage, G. E., Hollingshead, A. B., Drake, L. R., & Littlepage, A. M. (2008). Transactive memory and performance in work groups: Specificity, communication, ability differences, and work allocation. *Group Dynamics: Theory, Research, and Practice, 12*(3), 223–241.

Marks, M. A., DeChurch, L. A., Mathieu, J. E., Panzer, F. J., & Alonso, A. (2005). Teamwork in multiteam systems. *Journal of Applied Psychology, 90*(5), 964–971.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *The Academy of Management Review, 26*(3), 356–376.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology, 85*(2), 273–283.

Mathieu, J. E., Marks, M. A., & Zaccaro, S. J. (2001). Multiteam systems. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology* (Vol. 2, pp. 289–313). London: Sage.

Mathieu, J. E., Maynard, M. T., Taylor, S. R., Gilson, L. L., & Ruddy, T. M. (2007). An examination of the effects of organizational district and team contexts on team processes and performance: A meso-mediational model. *Journal of Organizational Behavior, 28*(7), 891–910.

McGrath, J. E. (1984). *Groups: Interaction and performance*. Englewood Cliffs, NJ: Prentice-Hall.

Mesmer-Magnus, J. R., & DeChurch, L. A. (2009). Information sharing and team performance: A meta-analysis. *Journal of Applied Psychology, 94*(2), 535–546.

Moreland, R. L. (1999). Transactive memory: Learning who knows what in work groups and organizations. In L. Thompson, J. Levine, & D. Messick, (Eds.). *Shared cognition in organizations: The management of knowledge* (pp. 3–31). Hillsdale, NJ: Erlbaum.

Moreland, R. L., Argote, L., & Krishnan, R. (1996). Socially shared cognition at work: Transactive memory and group performance. In J. L. Nye & A. M. Brower (Eds.), *What's social about social cognition? Research on socially shared cognition in small groups* (pp. 57–84). Thousand Oaks, CA: Sage.

Mullen, B., & Cooper, C. (1994). The relationship between group cohesiveness and performance: An integration. *Psychological Bulletin, 115*(2), 210–227.

Murase, T., Carter, D. R., DeChurch, L. A., & Marks, M. A. (2014). Mind the gap: The role of leadership in multiteam system collective cognition. *Leadership Quarterly, 25*(5), 972–986.

Salas E., Rosen M. A., Burke C. S., & Goodwin, G. F. (2009). The wisdom of collectives in organizations: An update of the teamwork competencies. In E. Salas, G. F. Goodwin, & C. S. Burke (Eds.), *Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches* (pp. 39–79). New York: Routledge.

Stajkovic, A. D., Lee, D., & Nyberg, A. J. (2009). Collective efficacy, group potency, and group performance: Meta-analysis of their relationships, and a test of a mediation model. *Journal of Applied Psychology, 94*(3), 814–828.

Tesluk, P., Mathieu, J. E., Zaccaro, S. J., & Marks, M. (1997). Task and aggregation issues in the analysis and assessment of team performance. In M. T. Brannick, E. Salas, & C. W. Prince (Eds.), *Team performance assessment and measurement: Theory, methods, and applications* (pp. 197–224). New Jersey: Lawrence Erlbaum Associates.

Thompson, J. D. (1967). *Organizations in action*. Chicago: McGraw-Hill.

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations*. Research Report 41-13. Princeton, NJ: Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/report/2013/jrps

Zaccaro, S. J., Marks, M. A., & DeChurch, L. A. (2012). Multiteam systems: An introduction. In S. J. Zaccaro, M. A. Marks, & L. A. DeChurch (Eds.), *Multiteam systems: An organization form for dynamic and complex environments* (pp. 3–31). New York: Routledge.

Chapter 4
Innovation in Team Interaction: New Methods for Assessing Collaboration Between Brains and Bodies Using a Multi-level Framework

Stephen M. Fiore and Katelynn A. Kapalo

Abstract As research on teams becomes increasingly sophisticated, scientists face challenges related to understanding collaboration at multiple levels of analysis, beyond that of the individual or the group alone. Grounded in Hackman's work on interaction and levels of analysis, this chapter explores theory development for understanding team collaboration from multiple perspectives. We argue that to enhance and improve the study of collaboration and to increase explanatory power, the development of theory must focus not only on the major issues at each level, micro, meso, macro, but also issues that cross these levels of analysis in team interaction. This method of cross-level analysis provides insight on some of the causal factors related to better understanding collaboration effectiveness. Furthermore, this chapter explores the need to leverage complementarity within and between disciplines to enhance our understanding of team interaction and to provide a more holistic method for assessing collaboration in a variety of complex domains.

Keywords Collaboration · Team interaction · Problem solving · Team science · Cross-level analysis · Micro · Meso · Macro

4.1 Teams and Technology: New Methods for Assessing Interaction and Collaboration Between Brains and Bodies

Over 400 years ago, a Dutch tinkerer named Zacharias Janssen, who worked in the fledgling spectacle industry, created a new tool. By engineering a set of lenses in a particular configuration, light could be manipulated such that objects could be

S.M. Fiore (✉) · K.A. Kapalo
University of Central Florida, 3100 Technology Parkway, Suite 140,
Orlando, FL 32826, USA
e-mail: sfiore@ist.ucf.edu

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_4

magnified many more times than before (Masters, 2008). Although not immediately recognized as such, this tool would revolutionize much of science. Within a few decades, Marcello Malpighi, an enterprising physician and biologist in Bologna, used this new technology to identify the capillaries posited in an earlier theory of the circulation of blood. Soon, the scientists of the day began their own modifications to this new tool, called a microscope, making it more powerful and more usable (Masters, 2008). But improving this technology was not the goal; it was merely the means to a newly realized end, that is, the ability to investigate tissue components that could not be seen with the human eye. For what they had perceived as a hidden world, was now visible thanks to this powerful new instrument—a tool that would help them discover the many and varied layers of this world. They could now explore biological intricacies and interconnections across various levels. At the micro-level of analysis, cellular components were now visible. At the meso-level, interactions between these cellular components and how they interact with one another were illuminated. Finally, at the macro-level, the complex systems, functioning as a result of multiple cellular interactions across levels, could be understood. By peeling away layers of organisms, subjecting them to forms of analysis never before possible, and studying inter-connections within and across these layers, they were able to observe and understand the beauty and the complexity of biological systems.

This brief tour of science history is merely an illustration, albeit a powerful one, of how a technology can revolutionize our understanding of the world around us. We are seeing a similar revolution in the study of collaboration. For, in research on groups and teams, we are having introduced to us, not just one, but many new tools and technologies helping us instrument and/or observe the world of interaction in ways never before possible. Importantly, though, we are observing interaction not just within, but also across, multiple levels, From this, we now have the opportunity to integrate levels of interaction in a meaningful way, and study collaboration in a variety of domains.

Within a volume emphasizing the importance of developing effective measures of collaboration via consideration of assessment approaches from a variety of disciplines, we submit that scientists must have an appropriate conceptual scaffold for understanding multiple forms and levels of analysis. This requires methods for diagnosing causal factors associated with collaboration effectiveness. In particular, by moving our analysis either one level up, or one level down, we can emphasize differing factors associated with teamwork. First introduced by Hackman (2003), the idea of shifting focus from an isolated level to a higher or lower level can lead to new insights into causal mechanisms that shape team process and performance. More importantly, bracketing a phenomenon of interest, via a level above and a level below, can increase the precision of explanation in that the "explanatory power of bracketing lies in crossing levels of analysis, not blurring them" (Hackman, 2003, p. 919). We build upon this to suggest that the simultaneous consideration of micro, meso, and macro-levels of collaboration, in addition to bracketing phenomena, can provide a rich explanatory framework for assessment.

From this, a truly multi-level theoretical perspective, that can specify constructs cutting across levels is within our reach (see Dansereau & Yamarino, 2002; Fiore et al., 2012).

4.2 The Context for Collaborative Assessment

In this chapter, we illustrate how multiple levels of analyses are moving us in important new directions for assessing collaboration. This provides grounding for a discussion of how integration of measures can be of value in the assessment of collaborative problem solving. We structure this summary by the level of analysis being used—micro, meso and macro-levels. First, we discuss recent research within these levels, on the study of collaboration. We then provide examples of how to integrate these to understand cross-level phenomena. Finally, we describe how such methods of assessment can be used to enrich our understanding of collaborative problem solving. We do this with the specific example of scientific problem solving as engaged by teams. In sum, we show that developments across disciplines are creating new methods for assessing interactions at the level of the brain, body, behavior, and network. Our goal is to help collaborative problem solving assessment researchers make sense of the varied studies emerging by more systematically considering the level of analysis in which collaboration is being studied so as to consider how to supplement more traditional forms of problem solving assessment.

4.2.1 Looking at Levels

Traditionally, team research focuses on a limited set of measures, and usually only at a single level of analysis. Although such approaches produce robust results, unidisciplinary assessment methods, and/or measures that too narrowly focus on one form of collaboration, or one level of analysis, can limit our understanding of the true richness of collaboration. As such, they do not adequately capture the complexity inherent in teamwork. Following calls for multi-level analyses (Dansereau & Yamarino, 2002; Hackman, 2003), we suggest that the assessment of collaboration match the complexity of team interaction by examining multiple levels and through a multi-method and multidisciplinary approach. In this way, we can address limitations in the literature on collaboration assessment.

Toward this end, we discuss multiple levels of analysis for analyzing concepts associated with collaboration and the developments being made in these areas. At the micro-level, we are interested in understanding the neurobiological and physiological *underpinnings* of social cognitive processes. Expanding outward, we move to the meso-level, encompassing *mediating* artifacts as well as movements and non-verbal behaviors between bodies. Finally, we reach the macro-level of analysis, which involves interactions *within and across* teams of teams and networks.

When we better understand concepts and methods for studying collaboration within levels, we can then move towards one of the more profound challenges in research on teams. This is creating and synthesizing theories and methods that can cross levels of analysis (cf. Hackman, 2003). With this, we can better understand the specific dynamics emerging in collaboration. To achieve this we must evolve team research into a truly interdisciplinary enterprise. Using this integrative approach, then, our goal is to help the field recognize the broader implications of interaction between bodies and brains and how this can be leveraged for more effective assessment of collaboration at all levels of analysis. For the purposes of this chapter, we discuss innovative assessments of collaboration and then relate these to collaborative problem solving as an specific form of collaboration.

4.2.2 Level One: Micro Level

As methods of assessment in neuroscience became more sophisticated and more robust, research has transitioned from a purely individual cognitive focus to understanding the biological mechanisms that drive social cognitive processes. The emerging area of social neuroscience solidified around these developments and brought about an important perspective on social cognitive mechanisms. Research at this more micro-level focuses on investigating the relationship between biological states, neurological properties, and collaboration.

Electroencephalogram (EEG) has matured into one of the important tools for research in the cognitive and neural sciences. EEG relies on electrodes attached to the scalp to detect electrical activity in the brain. Particular patterns of electrical impulses are used to assess varied forms of neural activity (e.g., attentional focus). Because of decreases in cost, and increases in reliability, EEGs are now one of the new ways for assessing neural activity in collaborative contexts.

To illustrate methods of collaboration assessment at this micro-level, EEG has been used to measure neural synchrony. This describes complementary or similar electrical impulses that emerge during collaborations. For example, in the context of coordination in body movement during a cooperative interaction, EEG was used in conjunction with motion tracking to study physiological changes in interacting pairs (Yun, Watanabe, & Shimojo, 2012). More specifically, phase synchrony was used to study inter-brain connectivity, the synchrony between the neurological responses of a dyad. Through this instrumentation, implicit interpersonal interactions were observable at a very fine-grain level based upon body movement synchronization. This study found that training in a cooperative task increased synchrony, “between cortical regions across the two brains [to suggest] that such inter-brain synchrony is a neural correlate of implicit interpersonal interaction” (Yun et al., 2012, p. 3). This illustrates how embodied approaches to assessing interaction can utilize methods developed within neuroscience. In particular, methods specifically assessing body movements, linked to neural assessment, can help us understand the relationship between interacting bodies and brains (cf. Valera, Thompson, & Rosch, 1991).

Synchrony in EEG activation has also been used during the complex coordinative process of guitar duets. This research expected brain areas associated with executive control and metacognition (the pre-frontal cortex, PFC) to be involved given the need to monitor teammates in the duet. This can be seen as a form of mental state attribution arising within the team while playing together. In this study, they examined coordination within guitar duets by recording EEG from each player in 12 duets (see Sanger et al., 2012). They assigned team roles for the duet by making one player a leader and the other a follower. Within-brain and between-brain coherence in time-frequency signals were then assessed. This study showed how synchronous oscillations in the duet varied dependent upon leader-follower assignments. Further, they found within-brain "phase locking" and between-brain "phase coherence" was heightened in the PFC when there were high demands placed on musical coordination. This can be interpreted as neural markers of interpersonal action coordination arising when there exists higher demands for monitoring teammates.

Body mirroring in collaboration, is another emerging area of research that continues to evolve. Research in this area examines joint action and biological function in the context of collaborative environments. Studies have demonstrated the influence of musical structure in choral singing on cardiovascular function by measuring the heart rate variability (HVR) and respiratory sinus arrhythmia (RSA) rates (Vickoff et al., 2013). This suggests that singing "as a group" can cause individual biological responses to synchronize.

Neuroendocrinology research is helping us understand how the neuropeptide oxytocin influences trust and cooperation in groups and can alter behaviors across groups (De Dreu, Shalvi, Greer, Van Kleef, & Handgraaf, 2012). Using a modification of the classic Prisoner's Dilemma game, this experiment studied the traditional patterns of interaction that can arise during game-play (e.g., reward or punishment). They found that oxytocin, when administered via nasal inhalation, influenced the desire to protect vulnerable group members. In other words, even when not personally threatened, oxytocin uptake produces prosocial behaviors, in this case, the desire to protect group members perceived as vulnerable (De Dreu et al., 2012). Such findings can help us understand micro-level methods to assess trust and motivation in terms of defensive capabilities that arise during collaboration.

Further, research has shown how neuropeptides change when team members engage in cooperative and collaborative behaviors. Levels of oxytocin were found to be related to group-serving tendencies during an incentivized poker game (Ten Velden et al., 2014). While De Dreu et al. (2012) outlined the effects of oxytocin towards vulnerable group members, Ten Velden et al. (2014) showed that participants decreased competitive behaviors when playing poker with an in-group member. Additionally, results indicated that participants receiving a dose of oxytocin were more likely to demonstrate cooperative behaviors when compared to the placebo group. This research suggests that, although oxytocin may not indiscriminately increase the prevalence of benevolence in humans, it may play a role in increasing cooperative behavior within groups.

These studies provide new insights on micro-level assessments by documenting that neurophysiological changes can be connected to interaction. This provides further support for using neuroscience in combination with traditional methods to measure collaborative interactions. As research advances in the study of the neurobiological underpinnings of behavior, we can use these to understand how they are related to traditional measures for studying collaboration. As we describe in more detail later, from this, then, we can consider how these related to the assessment of collaborative problem solving behaviors (e.g., heart rate variability and information sharing; oxytocin levels and back-up behaviors). As such, this can provide a more comprehensive picture and a richer understanding of interaction through assessments of neurological and biological markers of collaborative behavior.

4.2.3 Level Two: Meso Level

As we move beyond the neural level, we transition to what we label "meso-level" research, defined here as research focused on measuring interactions between bodies. This encompasses developments in the study of non-verbal behavior to offer rich insights from the observation of interactions. This also includes interactions, not just between team members, but also between members and artifacts in the world. These forms of external cognition are manipulated in service of shared information processing during collaborative problem solving (see Fiore & Schooler, 2004; Fiore et al., 2010). For example, research in human-computer interaction has blended psychological and computational approaches to examine how technologies are scaffolding group process and how artifacts and material objects mediate complex collaborative cognition.

At this meso-level, researchers have studied collaborative constructs such as shared awareness and common ground. For example, using a digital puzzle task that varied factors such as item complexity and visual feedback, research showed how shared visual spaces influence collaborative effectiveness (Gergle et al., 2013). This examined interactions in a problem solving task via study of "helpers," participants describing a puzzle configuration, and "workers," the participants actually assembling the puzzle. They found that visual spaces designed to scaffold the interaction, through the use of screens optimized for the task based on the role of the member in the dyad, influenced performance by altering conversational grounding and shared task awareness. This illustrates an important path for assessing cognition and communication in the context of material objects and how these relate to collaboration effectiveness.

Enhanced displays represent another important development for assessing how artifacts mediate interactions and cognition between interacting bodies. For example, in visual analytics, researchers have studied collaboration processes emerging during a complex task requiring distillation and comprehension of large amounts of information (Isenberg et al., 2012). Here, via study of mediated interaction through tabletop displays, researchers assessed collaboration patterns that arise when teams virtually manipulated hundreds of digital documents to solve problems requiring the integration of a vast amount of text. This provides insights for assessing the relationship between loosely and tightly coupled interactions "around" tasks, artifacts, and displays as team members collaborate to, for example, distill and synthesize information.

Developments within the field of "environmentally aware computing" are also allowing us to understand patterns of interaction related to any number of team outcomes. For example, by integrating the use of sociometric badges (i.e., wearable devices that collect social data such as proximity to, and amount of interaction with, others), with traditional surveys, research is studying the influence of collaboration and creativity (Tripathi & Burleson, 2012). This research assessed individual creativity but examined it in the context of team meetings via sociometric badges and the amount of interaction team members experienced. By studying interaction in situ, they developed a predictive model of creativity in teams in their organizational context. This study illustrates a powerful way to infuse new technology (sociometric badges) into traditional studies so as to improve assessment and gain a better understanding of collaboration embedded in context (see also Khan, this volume, Chap. 11).

Sensor technology is also providing new ways of assessing group performance in the actual context of interaction. Infrared optical systems and passive markers are now being used for kinematic data capture during group interaction (D'Ausilio et al., 2012). Here, non-verbal behavior was studied to examine movement patterns related to leadership in orchestras. This research was able to produce detailed computational analysis of the causal relations between a conductor's wand and violinists' elbow movement. From this, they were able to uncover trends in leadership that were then related to the aesthetic quality of music. This provides an unobtrusive method for assessing a complex form of interaction, that, when paired with appropriate analytic techniques, help us better understand traditional concepts like leader-follower behaviors as discussed in the collaboration literature.

In short, these studies illustrate how technologies are helping us study, at a finer-grain, and in new ways, the behavioral aspects of social interaction. At this meso-level we can directly observe patterns of movement associated with joint action as well as collaboration with artifacts in the environment. These provide insights into how team members monitor actions with each other and/or with cognitive artifacts to carry out collective goals. This moves us beyond a discussion of the biological bases of interaction, to a discussion of the bodily forms of interaction. Further, at this level, and with this technology, we can study how contextual factors are related to collaboration. We provide more specific detail later, but, in brief, by linking this with the micro-level, we can begin to envision how to integrate assessments of the neural underpinnings of collaboration with the behavioral interactions between team members to improve our understanding and assessment of collaborative problem solving in situ (e.g., EEG measures of engagement with task/system correlated with team process measures (cf. Stevens, Galloway, Wang, & Berka, 2012), this volume, Chap. 20; eye tracking with use of material artifacts during collaborative problem solving (cf. Olsen et al., Chap. 10).

4.2.4 Level Three: Macro Level

With the goal of understanding behavior and the influence of others on our interactions, we transition to the level with the broadest scope, the macro-level. This includes the study of teams of teams or large networks where subgroups emerge out of the interactions of hundreds, and sometimes thousands, or millions, of individuals. Developments in network science and social network analysis help us study these broad patterns of interaction across multiple time scales.

As an example, macro-level analyses using bibliometrics are providing new ways for understanding collaboration as it occurs in the real world. In a study of 20 million patents and publications, over 50-years, researchers found that collaboration in science is on the rise and that teamwork in science is having an increasing impact on the production of knowledge (Jones et al., 2008; Wuchty et al., 2007). But this form of macro-level analyses can be even more fine-grained. For example, network analyses were used to study successful forms of interaction in complex teamwork environments. To illustrate, research on scientific teamwork produced analytic techniques that simultaneously took into account patterns of prior co-authorship coupled with analysis of citation overlap. In a study of over 1000 collaborative proposals, this was used to help determine team assembly as well as predict collaboration success in scientific teams (see Contractor, 2013, for a discussion). These studies provide insights into local interactions by studying broader patterns of collaboration across thousands of teams that unfold over long periods of time.

In sports, interaction networks are helping to assess the patterns of effective team performance. For example, in studying nearly 300,000 passes in professional soccer, using metrics such as network intensity (e.g., the passing rate), and network centrality (e.g., player dominance), high intensity and low centralization were related to more effective game play (Grund, 2012). In a study of over 12,000 video game production teams (with over 130,000 individuals), and over several years, network analyses helped uncover the factors contributing to development of games considered to be highly innovative (De Vaan, Stark, & Vedres, 2015). They found that the repertoire of skills acquired by individuals contributes to success if team members are stylistically different; that is, when individuals with differing skill sets leverage their strengths to collaborate more effectively. Specifically, when teams were found to have more diversity in these skills and styles, they were more likely to produce unique or distinctive games. These studies provide innovative approaches for understanding behavior but also point us towards new targets for assessment (e.g., collaborative competencies).

Social network analysis is also providing insights into performance within virtual settings, in the context of Massively Multi-player Online Games (MMOGs). With data collected over multiple months, over 7000 players, and millions of messages, factors such as alliances, trades, and cooperation were used to understand how teams accomplished goals (Wigand et al., 2012). When dealing with competition, network analyses documented that intensive communication and coordination enhanced team performance and that successful players were more likely to receive, than send, messages.

Others have also used social network analysis at this more macro-level to study how groups form in virtual worlds. For example, community detection algorithms were developed from interaction data (e.g., thousands of entries in chat rooms) to help understand the relationship between the type of interaction and group formation. Prior group membership, in this context, within guilds, was found to be most predictive of future membership. Additionally, network centrality was also shown to predict patterns of joining and be more important than member skill sets (see Alvari et al., 2014). Although these studies take place in virtual worlds, tracking behaviors of thousands of individuals, and over long periods of time, provide a window into collaboration not available using traditional laboratory studies.

In sum, these macro-level studies provide a level of understanding not attainable through analysis of neural pathways or behavioral observations. Further, they help us understand teamwork in both real and virtual worlds and across thousands of collaborating groups. By focusing on team dynamics at the macro-level, we can see the factors that contribute to successful interaction beyond an individual level and in high fidelity situations (e.g., sports teams, project production teams). While the work of neuroscientists and behavioral researchers is not to be overlooked, there is value in assessing teams beyond highly controlled lab studies. Specifically, by limiting our scope to only the micro or meso-levels of analysis, researchers overlook the value of understanding interaction more broadly. Further, network analyses provide a viable method for extracting factors that influence collaborative problem solving performance without interfering in the interactions or affecting the outcome of the interaction. This also has important implications given that studies at the neural (micro) level, and even behavioral (meso) level can be criticized for the potential influence of devices and methods in measuring the form or outcome of interactions. Thus, the predictive power of macro-level studies comes from both their scale and from their assessment of performance in situ.

4.3 Integrating Assessments Across Levels

Although looking within these levels is illuminating, we now turn considering the integration of levels. This requires a truly multi-level theoretical perspective where researchers assess collaborations at multiple levels in order to better specify how they are conceptualizing construct(s) that can cut across levels (see Dansereau & Yamarino, 2002; Fiore et al., 2012). Further, as noted earlier, shifting focus to a higher or a lower level can lead to new insights into causal mechanisms that shape team process and performance. As an analytical approach, bracketing the main phenomenon via a level above and a level below, can provide more precise explanations by specifying and crossing levels of analysis (Hackman, 2003). We similarly suggest that simultaneous consideration of micro-, meso-, and macro-levels of collaboration, in addition to bracketing phenomena, can provide a richer explanatory framework for understanding collaboration effectiveness.

To illustrate, research crossing what we would call the micro- and meso-levels is adding to our understanding of team cognition (Stevens et al., 2012). Research is demonstrating the utility of neurophysiological measures to augment our understanding of team process. In a simulated Submarine Piloting and Navigation (SPAN) task, temporal measures of engagement were mapped to team events. These measures tended to align with the frequency with which team members communicated with one another. This work in neurophysiological measures coupled with team communications, shows how to link the micro- and meso-levels, to improve and integrate novel and traditional methods for assessing collaboration (also see Stevens et al., Chap. 20, for further discussion of research on in situ assessment of collaboration). Others have discussed the value of what we consider crossing levels through the use of eye-tracking in collaborative tasks (Olsen, Ringenberg, Aleven, & Rummel, 2015). Using a “dual eye-tracking” paradigm, where eye gaze of collaborating teammates is used, this research examined how individual level gaze patterns are related to team level processes such as communication and learning outcomes (see also Olsen et al., this volume, Chap. 10). This work moves across these micro and meso levels by measuring joint visual attention in learning contexts. As such, researchers can collect data beyond the self-reporting procedures to study across levels where the individual interaction with their environment and other teammates plays a role in the outcome of learning sessions.

The aforementioned studies provide a direction for innovations in assessment. But our goal is to push the field towards more integration of assessment crossing levels. As such, to further illustrate the value of this way of pursuing research on teams, we use scientific problem solving as an example context for complex collaborative assessment that would benefit from a multi-level and multi-method assessment approach. Scientific teams are more the norm in research and development as the nature of the problems being studied is becoming increasingly more complex (Fiore, 2008; Hall et al., 2008; Stokols et al., 2008). Further, collaborative problem solving in science teams is not confined to a particular field as it is increasingly practiced within and across a variety of disciplines cutting across the physical, social, life/health and computational sciences (Asencio et al., 2012; Börner et al., 2010; Falk-Krzesinski et al., 2011; Olson & Olson, 2013). In this section, consideration of micro-, meso-, and macro-levels, and their interactions, can illuminate our understanding of collaborative problem solving in science.

When considering collaboration assessment via a multi-level lens, we must consider complementary approaches (Klein, Canella, & Tosi, 1999; Kozlowski & Klein, 2000). First, there can be assessment approaches envisioning how variables at higher levels might moderate the relations of variables at lower levels. In scientific collaboration, this could include how macro-level behaviors influence micro-level attitudes. In our science team example, this might be a macro-level factor, such as the data-sharing infrastructure across teams of teams as might occur with multi-university collaborations, and how this could have a downstream and proximal influence on a micro-level factor like team trust. Second, there can be models that examine how individual level factors shape higher level contexts. Continuing with our collaborative problem solving example of a science team, this kind of micro- to meso-level effect could occur when demographic factors (e.g., multidisciplinary team consisting of social scientists and life scientists) influences collaboration factors at the team level (e.g., coordination losses because of lack of shared knowledge across team members). Our point in describing complementary approaches is that, by not taking these into account, research in collaborative assessment of scientific problem solving might inaccurately specify the nature of relations of interest, or, they might even miss relationships entirely.

To ground the above distinctions in our micro-, meso-, macro-level framework, we next provide a set of specific examples to illustrate how integration of measures could be of value in the assessment of collaborative problem solving in science teams. First, micro- and meso-levels could be crossed such that we can study how neurophysiological indicators are related to broader interaction behaviors. As an example, research could examine how neural synchrony relates to the development of common ground in communications within teams. In a science team, this could be demonstrated by using EEG to assess patterns of synchrony while members work through hypotheses generation during proposal writing. Additionally, neuropeptides could be correlated with artifact construction and use. For example, higher levels of oxytocin might predict willingness to contribute to the development of material objects in the science team as they work on a proposal (e.g., drawings of a conceptual model).

Micro-level factors can also be connected to the more macro-level. For example, phase locking during initial interactions, as measured via EEG, might be indicative of later group formations. More specifically, it could be that science teams demonstrating greater phase locking during initial proposal meetings are more likely to continue and form teams who successfully complete or win a proposal. We can also envision how meso- and macro-levels of collaboration are related. Assessments studying broad patterns of collaborative science might be related to the degree of document sharing and/or idea integration at meso-levels. For example, analyses of proposal generation across entire fields, such as could be done using data from funding agencies, could be supplemented with follow-up methods that look at successful and unsuccessful proposals and how team interactions are related to behaviors like more openly sharing methods or findings within proposal writing teams.

In sum, we can improve explanatory power by using this cross-level assessment approach to better diagnose causal factors associated with collaboration effectiveness. By moving the analytical lens either one level up, or one level down, we may be able to shed new light on important factors associated with collaborative problem solving in science teams.

4.4 Conclusions

As the technological landscape evolves, so does our ability to study collaborative problem solving. And, although effective collaboration is our end goal, we need to recognize the importance of leveraging the complementary approaches found among different disciplines in order to optimize our processes and understanding. The methods of different disciplines can provide greater insight into the assessment of collaboration than those of any single discipline alone. Further, it seems that *understanding collaboration* from several levels of analysis provides its own *opportunity for collaboration*. In particular, theory building across levels presents the means through which researchers across disciplines can collaborate to develop robust methods of studying and assessing interaction and collaboration (cf. Cikara & Van Bavel, 2014). By encouraging a broader approach to existing research questions, we can use this collaboration to our advantage.

These levels, taken together and separately, can leverage our existing knowledge to ultimately design and build collaborative measures for better understanding and assessing collaborative skills. Using a multi-level approach, we can draw comparisons between these levels to better inform the design of educational assessment. We are not limited to measures at one isolated level; team members and students alike must integrate their own knowledge with the environment and with their other team members. Using the theoretical and empirical advances we have recently made in the educational domain requires a level of understanding from multiple domains: psychology, biology, neuroscience to name a few, and more importantly, effective assessments that can be deployed in the environment of the learner. Drawing from the tools of disciplines pursuing research on collaboration and the need to assess collaboration from a learning perspective, we can identify some intersections and pinpoint areas for further research if we focus on a multi-level approach.

In sum, the purpose of this chapter was to demonstrate how multiple levels of analysis can inform our understanding of collaboration and our ability to develop tools, methods, and novel approaches for assessing collaboration. Just as the microscope uncovered the hidden layers of biological systems, these technologies are revealing the complex inter-connections within and across social systems. What must be recognized, though, is that these technologies are helping us better understand the concepts and constructs and the theories we have already developed. That is, they are providing a new perspective on concepts such as coordination, or communication, or even cooperation and conflict. With this chapter, we hope to push the field forward so as to capitalize on these developments. To do this, groups and teams researchers need to broaden their own collaborations and share new methods and measures. Further, stronger ties with experts in psychometrics and assessment are an additional form of interdisciplinary collaboration necessary to enhance the accuracy and the precision of these new methods and technologies. Only then, can we begin to generate *new* constructs and concepts in groups and teams research. And, only then, can we reap the intellectual rewards that these technologies promise through the development of new theories that transcend disciplines and provide a fuller understanding of groups and teams.

Acknowledgements The writing of this chapter was partially supported by Grant SES-0915602 from the National Science Foundation and Grant NAKFICB10 from the National Academies of Science Keck Futures Initiative, both awarded to the first author. The views, opinions, and findings contained in this article are the authors and should not be construed as official or as reflecting the views of the University of Central Florida, the National Science Foundation, or the National Academies of Science.

References

Alvari, H., Lakkaraju, K., Sukthankar, G., & Whetzel, J. (2014). Predicting guild membership in massively multiplayer online games. In *Social computing, behavioral-cultural modeling and prediction* (pp. 215–222). Berlin: Springer International Publishing.

Asencio, R., Carter, D. R., DeChurch, L. A., Zaccaro, S. J., & Fiore, S. M. (2012). Charting a course for collaboration: A multiteam perspective. *Translational Behavioral Medicine, 2*(4), 487–494.

Börner, K., Contractor, N., Falk-Krzesinski, H. J., Fiore, S. M., Hall, K. L., Keyton, J., et al. (2010). A multi-level systems perspective for the science of team science. *Science Translational Medicine, 2*(49), 1–5.

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science, 9*(3), 245–274.

Contractor, N. S. (2013). Some assembly required: leveraging web science to understand and enable team assembly. *Philosophical Transactions of the Royal Society. Series a, Mathematical, Physical, and Engineering Sciences*.

D'Ausilio, A., Badino, L., Li, Y., Tokay, S., Craighero, L., Canto, R., … & Fadiga, L. (2012). Leadership in orchestra emerges from the causal relationships of movement kinematics. *PLoS ONE, 7*(5), e35757.

Dansereau, F., & Yamarino, F. (Eds.). (2002). *Research in multi-level issues*. Oxford: Elsevier Science Ltd.

De Dreu, C. K. W., Shalvi, S., Greer, L. L., Van Kleef, G. A., & Handgraaf, M. J. J. (2012). Oxytocin motivates non-cooperation in intergroup conflict to protect vulnerable in-group members. *PLoS ONE, 7*(11), e46751. doi:10.1371/journal.pone.0046751.

De Vaan, M., Stark, D., & Vedres, B. (2015). Game changer: The topology of creativity. *American Journal of Sociology, 120*(4), 1144–1194.

Falk-Krzesinski, H. J., Contractor, N. S., Fiore, S. M., Hall, K. L., Kane, C., Keyton, J., … & Trochim, W. (2011). Mapping a research agenda for the science of team science. *Research Evaluation, 20,* 143–156.

Fiore, S. M. (2008). Interdisciplinarity as teamwork: How the science of teams can inform team science. *Small Group Research, 39*(3), 251–277.

Fiore, S. M., & Schooler, J. W. (2004). Process mapping and shared cognition: Teamwork and the development of shared problem models. In E. Salas & S. M. Fiore (Eds.), *Team Cognition: Understanding the factors that drive process and performance* (pp. 133–152). Washington, DC: American Psychological Association.

Fiore, S. M., Rosen, M. A., Pavlas, D., & Jentsch, F. (2012). Conceptualizing cognition at multiple levels in support of training team cognitive readiness. In *Proceedings of 56th annual meeting of the human factors and ergonomics society* (pp. 448–452). Santa Monica, CA: Human Factors and Ergonomics Society.

Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., & Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors, 52*(2), 203–224.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2013). Using visual information for grounding and awareness in collaborative tasks. *Human-Computer Interaction, 28*(1), 1–39. doi:10.1080/07370024.2012.678246.

Grund, T. U. (2012). Network structure and team performance: The case of English premier league soccer teams. *Social Networks, 34*(4), 682–690. doi:10.1016/j.socnet.2012.08.004.

Hackman, J. R. (2003). Learning more from crossing levels: Evidence from airplanes, orchestras, and hospitals. *Journal of Organizational Behavior, 24,* 1–18.

Hall, K. L., Feng, A. X., Moser, R. P., Stokols, D., & Taylor, B. K. (2008). Moving the science of team science forward: Collaboration and creativity. *American Journal of Preventive Medicine, 35*(2S), 243–249.

Isenberg, P., Fisher, D., Paul, S. A., Ringel, M., Inkpen, M., Inkpen, K., & Czerwinski, M. (2012). Collaborative visual analytics around a tabletop display. *IEEE Transactions on Visualization and Computer Graphics, 18*(5), 689–702.

Jones, B., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science, 322,* 1259–1262.

Klein, K. J., Cannella, A., & Tosi, H. (1999). Multilevel theory: Challenges and contributions. *Academy of Management Review, 24,* 243–248.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco, CA: Jossey-Bass.

Masters, B. R. (2008). History of the optical microscope in cell biology and medicine. *Encyclopedia of Life Sciences (ELS).* Chichester: Wiley. doi:10.1002/9780470015902.a0003082

Olsen, J. K., Ringenberg, M., Aleven, V., & Rummel, N. (2015). Dual eye tracking as a tool to assess collaboration. In *ISLG 2015 fourth workshop on intelligent support for learning in groups* (pp. 25–30).

Olson, J. S., & Olson, G. M. (2013). Working together apart: Collaboration over the internet. *Synthesis Lectures on Human-Centered Informatics, 6*(5), 1–151.

Sänger, J., Müller, V., & Lindenberger, U. (2012). Intra- and interbrain synchronization and network properties when playing guitar in duets. *Frontiers in Human Neuroscience, 6,* 312. doi:10.3389/fnhum.2012.00312.

Stevens, R. H., Galloway, T. L., Wang, P., & Berka, C. (2012). Cognitive neurophysiologic synchronies what can they contribute to the study of teamwork? *Human Factors: The Journal of the Human Factors and Ergonomics Society, 54*(4), 489–502.

Stokols, D., Misra, S., Moser, R., Hall, K. L., & Taylor, B. (2008). The ecology of team science: Understanding contextual influences on transdisciplinary collaboration. *American Journal of Preventive Medicine, 35*(2), S96–S115.

Ten Velden, F. S., Baas, M., Shalvi, S., Kret, M. E., & De Dreu, C. K. (2014). Oxytocin differentially modulates compromise and competitive approach but not withdrawal to antagonists from own vs. rivaling other groups. *Brain Research, 1580,* 172–179.

Tripathi, P., & Burleson, W. (2012). Predicting creativity in the wild: Experience sample and sociometric modeling of teams. *Proceedings of CSCW '12 computer supported cooperative work*, February 11–15, 2012. WA, USA: Seattle.

Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.

Vickhoff, B., Malmgren, H., Åström, R., Nyberg, G., Engvall, M., Snygg, J., et al. (2013). Music structure determines heart rate variability of singers. *Frontiers in Psychology, 4,* 334.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039.

Wigand, R. T., Agarwal, N., Osesina, I., Hering, W., Korsgaard, A., Picot, A., & Drescher, M. (2012). Social network indices as performance predictors in a virtual organization. In *Proceedings of the 4th international conference on computational aspects of social networks (CASoN 2012), Sao Carlos, Brazil*, November 21–23, 2012. Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*, 1036–1039.

Yun, K., Watanabe, K., & Shimojo, S. (2012). Interpersonal body and neural synchronization as a marker of implicit social interaction. *Nature Scientific Reports, 2,* 959. doi:10.1038/srep00959.

Chapter 5
Assessing Collaborative Problem Solving Through Conversational Agents

Arthur C. Graesser, Nia Dowell, and Danielle Clewley

Abstract Communication is a core component of collaborative problem solving and its assessment. Advances in computational linguistics and discourse science have made it possible to analyze conversation on multiple levels of language and discourse in different educational settings. Most of these advances have focused on tutoring contexts in which a student and a tutor collaboratively solve problems, but there has also been some progress in analyzing conversations in small groups. Naturalistic patterns of collaboration in one-on-one tutoring and in small groups have also been compared with theoretically ideal patterns. Conversation-based assessment is currently being applied to measure various competencies, such as literacy, mathematics, science, reasoning, and collaborative problem solving. One conversation-based assessment approach is to design computerized *conversational agents* that interact with the human in natural language. This chapter reports research that uses one or more agents to assess human competencies while the humans and agents collaboratively solve problems or answer difficult questions. AutoTutor holds a collaborative dialogue in natural language and concurrently assesses student performance. The agent converses through a variety of dialogue moves: questions, short feedback, pumps for information, hints, prompts for specific words, corrections, assertions, summaries, and requests for summaries. *Trialogues* are conversations between the human and two computer agents that play different roles (e.g., peer, tutor, expert). Trialogues are being applied in both training and assessment contexts on particular skills and competencies. Agents are currently being developed at Educational Testing Service for assessments of individuals on various competencies, including the Programme for International Student Assessment 2015 assessment of collaborative problem solving.

A.C. Graesser (✉) · N. Dowell · D. Clewley
University of Memphis, Memphis, TN, USA
e-mail: graesser@memphis.edu

N. Dowell
e-mail: ndowell@memphis.edu

D. Clewley
e-mail: dnclwley@memphis.edu

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_5

Keywords Autotutor · Conversational agents · Cohesion · Coh-Metrix · Collaborative problem solving · Emotions · Feedback · Intelligent tutoring systems · PISA · Trialogues · Tutoring

5.1 Introduction

Collaboration requires communication between two or more individuals during the process of learning, problem solving, or performing routine tasks that require coordination. This chapter focuses on conversation-based assessment of a person's competencies in one-on-one tutoring and in small groups during collaborative problem solving. Our distinctive slant is the use of computer agents in these assessments. That is, the human who is assessed holds conversations with one or more computer agents during the process of collaboration, and assessments are derived from the conversations.

Conversational agents are personified computer entities that interact with the human in natural language. Embodied conversational agents include talking heads or full-bodied animated avatars that generate speech, actions, facial expressions, and gestures. Disembodied agents send text messages or spoken messages without any visual depiction. The agents relevant to this chapter are adaptive to the actions, language, and sometimes the emotions of the learner, as opposed to delivering rigidly orchestrated displays of language and action. Adaptive agents have been developed to serve as substitutes for humans who range in expertise from peers to subject matter experts that tutor the learners. Agents can guide the learner on what to do next; hold collaborative conversations; deliver didactic content; and model ideal behavior, strategies, reflections, and social interactions.

Agents also track the performance, knowledge, skills, and various psychological characteristics of the humans online during these conversational interactions. This formative assessment of the human is of course essential for any adaptive learning or assessment environment. Dozens, hundreds, or even thousands of observations per hour are collected in log files and serve as indicators of performance and diverse psychological attributes (Dede, 2015; D'Mello & Graesser, 2012; Shute & Ventura, 2013; Sottilare, Graesser, Hu, & Holden, 2013). The scores derived from the massive data feed into assessment measures that potentially can meet high standards of reliability and validity. This chapter identifies different types of scores and measures that we have collected in our agent-based learning and assessment environments. Some of these scores and measures target the collaborative problem solving (CPS) proficiencies of individual humans in their ability to collaborate with others during the process of group problem solving. It is beyond the scope of this chapter to measure the CPS characteristics of groups of individuals, but we will briefly describe how agents were used in Programme for International Student Assessment (PISA) 2015 assessments of CPS (see Graesser, Foltz et al., in press; Graesser, Forsyth, & Foltz, 2016; OECD, 2013).

Adaptive conversational agents have become increasingly popular in contemporary learning environments. Some examples that have successfully improved student learning are AutoTutor (Graesser et al. 2004, 2012; Nye, Graesser, & Hu, 2014), DeepTutor (Rus, D'Mello, Hu, & Graesser, 2013), GuruTutor (Olney et al. 2012), Betty's Brain (Biswas, Jeong, Kinnebrew, Sulcer, & Roscoe, 2010), iSTART (Jackson & McNamara, 2013), Crystal Island (Rowe, Shores, Mott, & Lester, 2010), Operation ARA (Halpern et al., 2012; Millis et al., 2011), and My Science Tutor (Ward et al., 2013). These systems have covered topics in STEM (physics, biology, computer literacy), reading comprehension, scientific reasoning, and other domains and skills. These environments have online formative assessment with scores that are computed immediately and stored in a student model. A student model is a database that records the performance, knowledge, skills, strategies, and noncognitive psychological attributes of the student continuously over time (Sottilare et al., 2013). The scores in the student model provide input to computational mechanisms that decide what problem to present next and what dialogue moves of the agent to express next. Researchers also perform data mining analyses and machine learning modeling offline to discover conversational interaction patterns that influence student learning, motivation, and emotions (D'Mello & Graesser, 2012; Forsyth et al., 2013; Rowe et al., 2010).

What conversation patterns and discourse moves should be implemented in these adaptive conversational agents? These decisions are guided by the following considerations:

1. *Discourse moves and patterns in conversations among humans*. Early versions of AutoTutor were based on detailed analyses of hundreds of hours of face-to-face interaction of typical tutors (Graesser, Person, & Magliano, 1995), whereas GuruTutor was influenced by analyses of 10 expert human tutors (Cade, Copeland, Person, & D'Mello, 2008). Dialogue moves have also been analyzed on dozens of groups of three to four students interacting with a human mentor during collaborative learning and problem solving (Morgan, Keshtkar, Duan, & Graesser, 2012).
2. *Theoretical models*. There are theoretical models of ideal tutoring strategies (Graesser, D'Mello, & Cade, 2011) and of successful CPS in small groups (Fiore et al., 2010; Graesser, Foltz et al., in press; OECD, 2013; Fiore, this volume; Salas & Reyes, this volume).
3. *Data mining discoveries*. Successful and unsuccessful conversation patterns can be discovered from machine learning analyses of log files (Forsyth et al., 2013; Rosé et al., 2008; Rosé, Howley, Wen, Yang, & Ferschke, this volume; He & M. von Davier, this volume).
4. *Technical limitations*. Some discourse moves and conversation patterns are very difficult or impossible to implement because of limitations of current computational linguistics techniques (Jurafsky & Martin, 2008). For example, computers cannot reliably interpret messages that have complex logical derivations or precise mathematical expressions. Computers cannot reliably generate

discourse moves that dynamically build on the content from multiple conversational turns.

The remainder of this chapter turns to the scoring of human contributions in some agent-based conversational systems. We start with dialogues, then move on to trialogues (two computer agents and a human). The primary focus will be on scoring collaborative problem-solving competencies, but some of these build on other cognitive or noncognitive attributes.

5.2 Conversational Dialogues

The simplest agent interaction is a dialogue in which the human interacts with only one agent. The agent can take on different roles (expert, tutor, peer), abilities (low to high knowledge), and collaborative style (cooperative, helpful, adversarial, unresponsive). AutoTutor is a tutor agent that attempts to get the human to actively contribute through verbal messages and actions, with the goal of promoting active student learning (Graesser et al., 2004, Graesser, Jeon, & Dufty, 2008, 2012; Nye et al., 2014). The students' contributions can be either typed or spoken, but most of the research has accepted student input because learning does not significantly differ for typed versus spoken contributions (D'Mello, Dowell, & Graesser, 2011). AutoTutor presents problems to solve and difficult questions that require reasoning, typically with one to five sentences in an ideal answer. The student and tutor co-construct a solution or answer by multiple conversational turns (Chi, Siler, Yamauchi, Jeong, & Hausmann, 2001; Graesser et al., 2008). AutoTutor's pedagogical objective is well justified because meta-analyses report effect sizes between $\sigma = 0.20$ and 1.00 when comparing human tutoring to classroom teaching and other suitable comparison conditions (Cohen, Kulik, & Kulik, 1982; Graesser et al., 2011; VanLehn, 2011). The learning gains of AutoTutor are approximately the same as those of human tutors (Graesser et al., 2008, 2012; Nye et al., 2014; VanLehn, 2011; VanLehn et al., 2007).

Both AutoTutor and human tutors follow a systematic conversational mechanism that is called *expectation and misconception-tailored dialogue* (Graesser et al., 2008, 2012). That is, tutors anticipate particular correct answers (called *expectations*) and particular *misconceptions* when they ask the students challenging questions (or problems) and track their reasoning. As the students articulate their answers over multiple conversational turns, the student contributions are compared with the expectations and misconceptions, and the tutor thereby forms an approximate model of the students' proficiency. More specifically, suppose that there are three expectations and two misconceptions associated with a problem. Semantic match scores (varying between 0 and 1) are computed between the expectations and (a) the student's contributions in a single turn (a local match score), (b) the student's contributions over multiple turns (a cumulative match score), and (c) contributions up through the final turn of the conversation (called the final match score). Such match scores are computed for each of the three expectations and two misconceptions over the course of a conversation that stretches as long as 100 turns. The profile of the five final match scores (E1, E2, E3, M1, M2) can be used as an assessment of how much the student contributed to the co-construction of the answer/solution; the tutor agent gets credited with filling in the remaining information that covers the three expectations. Therefore a student profile of (0.8, 0.9, 0.7, 0.1, 0.0) would be excellent because the student covered the expectations quite well and expressed very little about the misconceptions. In contrast, a student with a (0.0, 0.1, 0.2, 0.1, 0.0) profile would contribute very little relevant information, and a student with a (0.4, 0.5, 0.6, 0.8, 0.1) profile would have an M1 misconception.

The preceding scores require a computational assessment of semantic matching. Fortunately, advances in natural language processing research have made major progress in the accuracy of these semantic matches in computers. These semantic match algorithms have included keyword overlap scores, word overlap scores that place higher weight on lower frequency words in the English language, scores that consider the order of words, latent semantic analysis cosine values, regular expressions, and procedures that compute logical entailment (Cai et al., 2011; Graesser & McNamara, 2012; Rus, McCarthy, McNamara, & Graesser, 2008). These automated semantic match scores are nearly as reliable as human expert annotators and are computed immediately in AutoTutor. It is beyond the scope of this chapter, however, to describe these semantic match algorithms.

AutoTutor generates dialogue moves that encourage the students to generate content and eventually cover the expectations. AutoTutor generates *pumps* (e.g., tell me more, what else) early in the conversation for a main question/problem in an attempt to get the students to express what they know. This is important because most students express only one to two sentences in response to the main question/problem, even though they know much more. After this first step of information gathering through the main question/problem and pumps, the tutor focuses on each of the expectations (E1, E2, E3) that the student has not covered, one at a time, and generates dialogue moves to get the student to articulate E_i). For each uncovered expectation E_j, the tutor invokes a hint → prompt → assertion cycle: first a *hint*, then a *prompt* question to elicit an unexpressed word (if the student does not give a good answer to a hint), and then an assertion (if the student does not give a good answer to the prompt). Thus, there is a systematic way to score a student's contribution to covering expectation E_j. The student gets full 1.0 credit if the semantic match score meets or exceeds some threshold for E_j after the main question/problem is asked; the student gets 0.67 credit if the answer yields an above-threshold semantic match after the hint; 0.33 credit if the answer is above threshold after the prompt; and 0.00 if the student is subthreshold and the tutor ends up asserting E_j. The final score for the student's contribution would be [(S1 + S2 + S3) − (S4 + S5)]. These student profiles have been validated in AutoTutor studies that compare these scores with objective tests of the subject matter knowledge of the learners (Jackson & Graesser, 2006). It should be noted that misconceptions are immediately corrected by AutoTutor when expressed by the students and would be counted against the students' performance scores.

The pump-hint-prompt-assertion sequences are generated by AutoTutor to optimize extraction of whatever knowledge the student has on each expectation. These dialogue moves lead to a reasonable assessment of what the student knows. There are other dialogue moves of AutoTutor that might help students learn but might not be appropriate for assessment per se. For example, AutoTutor gives short feedback (positive, neutral, negative) after learner turns that contribute to the answer. It is appropriate to give positive and negative feedback to help the students learn, but this would bias the student in pure assessment conversations; instead, neutral feedback (e.g., okay, uh-huh) is appropriate for assessment. As another example, sometimes the student asks questions in a conversation. The frequency of student questions is surprisingly low to modest in most classroom and tutoring contexts (Graesser et al., 1995; Graesser, McNamara, & VanLehn, 2005), but they do periodically occur. The tutor would not answer questions in pure assessment contexts but rather would pass the ball to the student (e.g., How would you answer your question?).

There are other scores of the student's language and discourse that are tracked for assessment. AutoTutor segments the information in the student turns into speech acts and classifies the speech acts into different categories: questions, short responses (e.g., yes, okay), assertions, metacognitive expressions (I'm lost, now I understand), metacommunicative expressions (What did you say?), and expressive evaluations (This is frustrating, I hate this material). The proportions of student contributions classified in these different categories are diagnostic of self-regulated learning, as in the case of student questions (Graesser et al., 2005). The frequency or proportion of student contributions that are [assertions + questions] is a reasonable index of the extent to which a student takes initiative in a conversation. The student assertions can be analyzed on many dimensions of meaning in addition to semantic overlap with expectations and misconceptions. For example, AutoTutor analyzes the assertions on vagueness, relevance to the subject matter, newness (i.e., adding new information to the conversation), and verbosity (number of words). These dimensions of student language are to some extent diagnostic of the learner's emotions (D'Mello & Graesser, 2012), such as frustration, confusion, and boredom, but it is beyond the scope of this chapter to discuss assessment of student emotions. The assertions of students have also been analyzed by other computational linguistics tools, such as Coh-Metrix (Dowell et al., 2014; McNamara et al., 2014) and Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007). For example, Coh-Metrix assesses the extent to which the learner uses formal language (with abstract words, complex syntax, high cohesion, and an informational genre) versus conversational language (with concrete words, simple syntax, low cohesion, and narrative genre). All of these dimensions of language and discourse can be automatically tracked in assessments of CPS in two-party conversations.

5.3 Conversational Trialogues

Our contention is that adding a second agent to form a *trialogue* will have intriguing benefits in improving both learning and assessment (Graesser, Forsyth, & Lehman, in press; Graesser, Li, & Forsyth, 2014; Graesser, McNamara, Cai, Conley, Li, & Pennebaker, 2014). In addition to AutoTutor, multiple agents have been incorporated in many learning environments with agents, such as Betty's Brain (Biswas et al., 2010), iSTART (Jackson & McNamara, 2013), and Operation ARIES! (Forsyth et al., 2013; Halpern et al., 2012; Millis et al., 2011). Multiple agents have also been implemented in assessment environments with trialogues (Zapata-Rivera, Jackson, & Katz, 2015) and *tetralogues* (two humans with two agents; Hoa et al. this volume).

Researchers have recently been exploring several configurations of trialogues to better understand how trialogues can be productively implemented for particular students, subject matters, depths of learning, and assessments (Cai, Feng, Baer, & Graesser, 2014; Graesser, Forsyth, & Lehman, in press; Graesser, Li, & Forsyth, 2014; Zapata-Rivera, Jackson, & Katz, 2015). For example, the trialogue designs in Table 5.1 have relevance to assessment of CPS in addition to learning.

The six trialogue designs in Table 5.1 do not exhaust the design space, but they do illustrate how assessment can be accomplished for different populations, competencies, and ranges of human abilities. Consider some examples. AutoTutor

Table 5.1 Trialogue designs in learning and collaborative problem solving

No.	Design	Description
1	Vicarious observation with limited human participation	Two agents communicate with each other and exhibit social interaction, answers to questions, problem solving, or reasoning. The two agents can be peers, experts, or a mixture. The two agents occasionally turn to the human and ask a prompt question (inviting a yes/no or single word answer) that can be easily assessed automatically. This trialogue design is appropriate for beginning phases of instruction, individuals with low knowledge, and those who have difficulty expressing their ideas verbally
2	Human interacts with two peer agents that vary in proficiency	The peer agents can vary in knowledge and skills. In assessment contexts, the computer can track whether the human is responsive to the peer agents, correctly answers peer questions, corrects an incorrect contribution by a peer, and takes initiative in guiding the exchange

(continued)

Table 5.1 (continued)

No.	Design	Description
3	Expert agent staging a competition between the human and a peer agent	There is a competitive game (with score points) between the human and peer agent, with the competition guided by the expert agent. A competitive game can be motivating for individuals in both learning and assessment environments
4	Expert agent interacting with the human and peer agent	There is an exchange between the expert agent and the human, but the peer agent periodically contributes and receives feedback. Negative short feedback can be given to the peer agent on bad answers (the agent takes the heat), whereas similar answers by the human receive neutral feedback. This circumvents direct negative feedback to the human, which can discourage some individuals
5	Human teaches/helps a peer agent with facilitation from the expert agent	This human teaches or helps the peer agent in need, with the expert agent rescuing problematic interactions. This trialogue design is suited to knowledgeable and skilled humans who have the ability to take charge of the interaction and solve the problem
6	Human interacts with two agents expressing false information, contradictions, arguments, or different views	The discrepancies between agents stimulate cognitive disequilibrium, confusion, and potentially deeper learning. There can be a focus on subtle distinctions that often are important for assessment

trialogues have been used in the Center for the Study of Adult Literacy (CSAL) to help readers 16 years and older develop and improve their comprehension skills (Graesser, Baer et al., 2015). These adults not only have difficulty reading but also have very low writing skills, confidence, and self-efficacy. When decisions were made on creating 35 learning lessons (30 min per lesson on average), we tended to select trialogue designs 1, 3, and 4 in Table 5.1 to optimize motivation and self-efficacy. In contrast, the trialogues in Operation ARIES! were designed to help college students learn research methods and apply principles of science in their reasoning, such as correlation does not imply causation and comparison groups are needed in experimental designs (Millis et al., 2011). Trialogue designs 2, 4, 5, and 6 in Table 5.1 were most dominant for these more knowledgeable and capable individuals. In some learning phases, the selection of trialogue design was adaptive to the knowledge of the student; the students who were not performing well received trialogue design 1, whereas the more knowledgeable students received designs 4 and 5. Trialogue design 6 was implemented in a series of studies that had college students read case studies that exhibited bad science, followed by conversational trialogues that critiqued the studies on violating particular science principles (D'Mello et al., 2014; Lehman et al., 2013). These studies showed that deeper learning was achieved for students who experienced cognitive disequilibrium and confusion when the two agents contradicted each other or expressed false information.

The trialogue designs have been implemented in complex computer environments that allow the humans to express themselves in different ways. The environments have ranged from interactions in chat rooms to interactions with multimedia and virtual reality (Cai et al., 2014; Zapata-Rivera et al., 2015). The communication channels of agents have varied from disembodied chat messages to full-bodied avatars in the virtual words. It is important to consider the alternative forms of human input in these trialogue designs. The easiest input is to perform a simple action, such as to click (touch) an option on the computer interface or a multiple-choice item. Instead of the human typing in open-ended verbal responses, he or she can make a selection from a presented set of three to five chat options. This approach was desired in CSAL AutoTutor because the adults had major difficulties in writing. It was also pursued in the PISA 2015 assessment of CPS because of logistical constraints, as described shortly. Limiting the human to three to five chat options indeed has a number of desirable features from the standpoint of assessment. The options can be stacked to focus on particular assessment constructs and subtle discriminations. The options allow easy automated scoring with unambiguous performance assessment. The potential downside is that humans sometimes want to express something in a conversational turn that is not listed among the chat options. Automated scoring of open-ended natural language is a reasonable approach given the advances in computational linguistics and discourse science, as was discussed in the section on conversational dialogue. It is possible to compute semantic match scores to expectations (e.g., to assess the correctness of their responses) and to classify human verbal contributions into different categories of speech acts (e.g., to assess how much initiative the human is taking in the conversation). All of these forms of human input can be objectively and automatically scored, as we discussed in the conversational dialogue section.

Agent-based assessments with trialogues allow considerable control over the conversations, especially compared with three-party conversations among humans. At the same time, it is important to acknowledge that the design of trialogues has additional computational challenges over dialogues because of the added complexity of the three-party conversations compared with the two-party dialogues. However, the complexity can be managed by thinking of a trialogue as functionally a dialogue between the human and a coordinated pair of agents plus media activities. More specifically, the discourse contributions of the two agents (A1, A2) and the media (M) can be coordinated so that each [A1, A2, M] sequential display is functionally a single episodic unit (U) that the human responds to through language, action, or silence in a particular human turn (HT).

A more technical discussion may help clarify how complexity can be managed with some coordination. There is a finite-state transition network that alternates episodic units (U) and HTs, analogous to a dialogue. That is, there is a small number of episodic unit states and a small number of states of HTs, with transitions specifying which states lead to other states. There can be conditional branching in the state transition network (STN) so that the computer's generation of episodic unit U_i at turn $n + 1$ is contingent on the state of the human turn HT_j at turn n. There is a small number of states associated with each HT_j, such as correct, incomplete, incorrect, or no response. The complexity of the branching depends on the number of finite states. Moreover, assessments normally require a specific set of episodic units that all test takers experience. Consequently, the STN in most agent-based assessments normally has a set of fixed episodic units (U_1, U_2, … U_m) distributed throughout the conversation (called convergence zones); scores are computed on the HTs that immediate follow each fixed episodic unit or the set of HTs that occur between the fixed episodic unit and the next fixed episodic unit. There is an exchange (oscillation) between the fixed episodic units and forest of conditional branching paths between the fixed episodic units. The complexity of the STNs can become quite rich, so authoring tools have been developed with chat maps and other visualization techniques to assist the content developer (Cai, Graesser, & Hu, 2015; Zapata-Rivera et al., 2015).

Given this STN formulation of trialogues, it is possible to score performance in ways analogous to conversational dialogues. For example, the agents can generate pump-hint-prompt-assertion cycles just as was expressed for tutoring, with the human receiving scores of 1.00, 0.67, 0.33, and 0 for correct responses before/after the pump, after the hint, after the prompt, versus after the agent's assertion, respectively. The occurrence of a hint, prompt, and assertion is contingent on the human's performance at each prior step in the pump-hint-prompt-assertion cycle. The [A1, A2, M] units would of course have to specify whether A1, A2, or M was set up as an initiator for the human to respond to. For open-ended verbal responses, there needs to be criteria on how good the verbal response needs to be before it is classified as a good answer versus other categories of answers. Alternative paths need to be set up that emanate from each category of human response. This includes human response categories of silence (no response), partial answer, off-topic, vague, metacognitive (I'm lost), and so on. Each of these categories needs to have follow-up episodic units or paths of units in the STN until the next fixed episodic unit occurs. Scores for a fixed unit U_i need to be computed for each of these paths. To keep the assessment manageable, it is wise to have the agents express *rescue* moves (asserting the correct answer and saying "let's move on") to cut off the mini-conversation that is launched after U_i and to start the next fixed episodic unit. Once these scores are specified for each fixed unit and the associated paths in the STN after each unit, it is possible to compute an overall score for the entire conversation by integrating over all of the fixed episodic units. Scores can also be broken down for particular skills, such as correctness, verbosity, initiative, and responsiveness.

5.4 Assessment of Collaborative Problem Solving in Programme for International Student Assessment 2015

The assessment of CPS for PISA 2015 (Graesser, Foltz et al., 2015, Graesser, Forsyth, & Foltz, 2016; OECD, 2013) has adopted the preceding approach of using agents, fixed episodic units, STNs with chat maps, a small set of chat alternatives at HTs, and a similar scoring methodology. The fixed units with a small number of chat alternatives provide data that are analogous to the multiple-choice tests that are familiar to the psychometric communities. One can apply the normal item response theory models with either dichotomous or polytomous scoring.

The following definition of CPS was articulated in the PISA 2015 framework for CPS:

> Collaborative problem solving competency is the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution.

Interestingly, the unit of analysis for the competency is the individual in a group rather than the group as a whole. The competency is an assessment on how well the individual interacts with agents during the course of problem solving. The overall competency is based on: (1) establishing and maintaining shared understanding, (2) taking appropriate action, and (3) establishing and maintaining team organization. This competency is crossed with different stages of problem solving that were adopted in the PISA 2012 assessment of individual complex problem solving: (A) exploring and understanding, (B) representing and formulating, (C) planning and executing, and (D) monitoring and reflecting. There were expected achievements in each of the 12 skills in the resulting 3 × 4 matrix, and each of the skills was scored. The scores were determined by the human's actions and chat selections in the fixed episodic units.

An adequate assessment required a judicious creation of tasks, group compositions, fixed episodic units, action options, and chat options. The assessment had to be completed in two 30-min sessions for any one individual, so efficiency was essential. The individual would be exposed to four to six problem-solving tasks with a diverse profile of agents (e.g., agreeable–disagreeable, cooperative–defiant, helpful–useless, responsive–unresponsive, correct–incorrect). Several dozen countries were assessed on CPS, so open-ended human responses were impossible. The selection of chat alternatives was critical so that the correct alternative was not obvious and the distractors reflected meaningful constructs. The correct alternatives could not be correlated with superficial features, such as politeness, taking charge, or specificity of actions. The superficial features of visual and spoken persona were also problematic, so the PISA assessments had disembodied chat messages.

It is important to acknowledge that these logistical constraints could not be achieved by an assessment environment among a group of humans without agents.

There is no control over what human partners in a conversation will say and do. The score for a particular individual in a conversation depends on group partners, so the scores for the individual will presumably be sensitive to whether the partners are social loafers or leaders. The individual human needs to be put in a number of groups with a number of partner characteristics before a sensible score can be achieved for that individual. Once again, the goal of PISA 2015 CPS was to assess the CPS proficiencies of an individual interacting in a group, not the group performance as a whole!

Some obvious questions arise about the validity of the agents in CPS assessment compared to human interaction. This is of course an empirical question. However, it is difficult to imagine that a meaningful assessment can emerge from open-ended communication among humans as they solve problems as a group within a limited amount of time. It takes time to hold a conversation among new humans, so there are worries about the time constraints within an assessment that lasts 1 h. There is no guarantee that a particular person is assigned to other humans that represent a broad profile of abilities and conversational styles, so that would compromise the validity of the person's assessment. There is no practical method of assembling complex group compositions in synchronous computer-mediated communication, so there is a low likelihood that three to four individuals in some groups could be assembled. Nevertheless, it would be worthwhile as a research question to assess how well human-to-human versus human-to-agent can be assessed on the 12 cells in the 4 × 3 matrix on PISA's theoretical framework on CPS.

5.5 Conclusions

This chapter has described how computer agents can be used in the assessment of collaboration. We have identified conversation patterns that provide conversation-based assessment, some scoring methods, some differences between dialogues and trialogues, and some challenges in assessing CPS. The role of agents in PISA 2015 CPS is also defended in the context of a large set of logistical constraints. The constraints include limited assessment time, a scientifically created sample of group composition, the prudent selection of tasks that expose skills in the CPS theoretical framework, and limitations of school schedules and computer networking.

The elephant in the room continues to be the question of how conversation-based assessments with computer agents compare to human-to-human collaborations on the same problem-solving tasks. The tasks would of course need to be the same in AH (agents with human) and HH (human with humans). Otherwise, the comparison is noncommensurate and invalid. The time constraints and the theoretical CPS framework would need to be the same as well. In the PISA assessments, there would need to be multiple languages and cultures as well as the need to assess performance quickly and economically within budget considerations.

There have been serious attempts to assess CPS among humans without any agents (Care & Griffin, 2014; Griffin & Care, 2015). In these efforts, computer-mediated CPS has been annotated on discourse categories with an eye to automation for CPS in 21st Century Skills. Unfortunately, progress in computational linguistics and discourse science is not at the point where these annotations can be reliably generated automatically. Also, establishment of interjudge agreement on coding among humans is a persistent challenge in achieving adequate reliability as well as minimizing time and expense. We do believe these efforts are extremely important but not ready for assessments of CPS processes and outcomes in the near future for multiple languages and cultures.

Meanwhile, we argue that the most prudent approach to assessing CPS for the next decade is to pursue the world of conversational agents. At one extreme, item developers can orchestrate a principled, rich display of media, agent speech acts, and interacting agents, with minimal input from the person being assessed (e.g., a click, scroll, drag and drop, word, or phrase). At the other extreme, item developers can create a task that requires conversation in natural language, facial expressions, gestures, actions, and other input modalities. Our view is that the first option is more pragmatic than the second option, but research on the second option is worthwhile as a long-term goal.

Acknowledgements The research was supported by the National Science Foundation (grants DRK-12-0918409, DRK-12 1418288, and DIBBS-1443068), the Institute of Education Sciences (grants R305A 130030, R305A100875, and R305C120001), the Army Research Laboratory (contract W911NF-12-2-0030), and the Office of Naval Research (contracts ONR N00014-12-C-0643). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, or DoD.

References

Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology-Enhanced Learning, 5,* 123–152.

Cade, W., Copeland, J., Person, N., & D'Mello, S. K. (2008). Dialogue modes in expert tutoring. In B. Woolf, E. Aimeur, R. Nkambou & S. Lajoie (Eds.), *Proceedings of the ninth international conference on intelligent tutoring systems* (pp. 470–479). Berlin: Springer.

Cai, Z., Feng, S., Baer, W., & Graesser, A. (2014). Instructional strategies in trialog-based intelligent tutoring systems. In R. Sottilare, A. C. Graesser, X. Hu & B. Goldberg (Eds.), *Design recommendations for intelligent tutoring systems: Adaptive instructional strategies* (Vol. 2, pp. 225–235). Orlando, FL: Army Research Laboratory.

Cai, Z., Graesser, A. C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., & Butler, H. (2011). Trialog in ARIES: User input assessment in an intelligent tutoring system. In W. Chen & S. Li (Eds.), *Proceedings of the 3rd IEEE international conference on intelligent computing and intelligent systems* (pp. 429–433). Guangzhou, China: IEEE Press.

Cai, Z., Graesser, A. C., & Hu, X. (2015). ASAT: AutoTutor script authoring tool. In. R. Sottilare, A. C. Graesser, X. Hu & K. Brawner (Eds.), *Design recommendations for intelligent tutoring systems: Authoring tools* (Vol. 3, pp. 199–210). Orlando, FL: Army Research Laboratory.

Care, E., & Griffin, P. (2014). An approach to assessment of collaborative problem solving. *Research and Practice in Technology Enhanced Learning, 9,* 367–388.

Chi, M. T. H., Siler, S., Yamauchi, T., Jeong, H., & Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science, 25,* 471–534.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19,* 237–248.

D'Mello, S., Dowell, N., & Graesser, A. C. (2011). Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology: Applied, 17,* 1–17.

D'Mello, S., & Graesser, A. C. (2012). Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies, 5,* 304–317.

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. C. (2014). Confusion can be beneficial for learning. *Learning and Instruction, 29,* 153–170.

Dede, C. (2015). *Data-intensive research in education: Current work and next steps*. Retrieved from http://cra.org/cra-releases-report-on-data-intensive-research-in-education/

Dowell, N. M., Graesser, A. C., & Cai, Z. (in press). Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics., XX*, XX–XX.

Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., & Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors, 52,* 203–224.

Forsyth, C. M., Graesser, A. C., Pavlik, P., Cai, Z., Butler, H., Halpern, D. F., et al. (2013). OperationARIES! Methods, mystery and mixed models: Discourse features predict affect in a serious game. *Journal of Educational Data Mining, 5,* 147–189.

Graesser, A. C., Baer, W., Feng, S., Walker, B., Clewley, D., Hayes, D. P., & Greenberg, D. (2015). Emotions in adaptive computer technologies for adults improving learning. In S. Tettegah & M. Gartmeier (Eds.), *Emotions, technology, design and learning: Communication, for with and through digital media* (pp. 1–35). San Diego, CA: Elsevier.

Graesser, A. C., D'Mello, S. K., & Cade, W. (2011). Instruction based on tutoring. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 408–426). New York: Routledge Press.

Graesser, A. C., D'Mello, S. K., Hu, X., Cai, Z., Olney, A., & Morgan, B. (2012). AutoTutor. In P. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation, and resolution* (pp. 169–187). Hershey, PA: IGI Global.

Graesser, A. C., Foltz, P. W., Rosen, Y., Shaffer, D. W., Forsyth, C., & Germany, M. (in press). Challenges of assessing collaborative problem solving. In E. Care, P. Griffin & M. Wilson (Eds.), *Assessment and teaching of 21st century skills*. Heidelberg, Germany: Springer.

Graesser, A. C., Forsyth, C. M., & Foltz, P. (2016). Assessing conversation quality, reasoning, and problem solving performance with computer agents. In B. Csapo, J. Funke & A. Schleicher (Eds.), *On the nature of problem solving: A look behind PISA 2012 problem solving assessment* (pp. 275–297). Heidelberg, Germany: OECD Series.

Graesser, A. C., Forsyth, C., & Lehman, B. (in press). Two heads are better than one: Learning from computer agents in conversational trialogues. *Teachers College Record.*

Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes*, *45*, 298–322.

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science*, *23*, 374–380.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *Elementary School Journal, 115,* 210–229.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers, 36,* 180–193.

Graesser, A. C., & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 307–325). Washington, DC: American Psychological Association.

Graesser, A. C., McNamara, D. S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point & Query, AutoTutor, and iSTART. *Educational Psychologist, 40,* 225–234.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9,* 495–522.

Griffin, P., & Care, E. (Eds.). (2015). *Assessment and teaching of 21st century skills: Methods and approach*. Dordrecht, Netherlands: Springer.

Halpern, D. F., Millis, K., Graesser, A. C., Butler, H., Forsyth, C., & Cai, Z. (2012). Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. *Thinking Skills and Creativity, 7,* 93–100.

Jackson, G. T., & Graesser, A. C. (2006). Applications of human tutorial dialog in AutoTutor: An intelligent tutoring system. *Revista Signos, 39,* 31–48.

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology, 105,* 1036–1049.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Lehman, B., D'Mello, S. K., Strain, A., Mills, C., Gross, M., Dobbins, A., & Graesser, A. C. (2013). Inducing and tracking confusion with contradictions during complex learning. *International Journal of Artificial Intelligence in Education, 22*, 85–105.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, MA: Cambridge University Press.

Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A., & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou & J. Lakhmi (Eds.), *Serious games and edutainment applications* (pp. 169–196). London, England: Springer.

Morgan, B., Keshtkar, F., Duan, Y., & Graesser, A. C. (2012). Using state transition networks to analyze multi-party conversations in a serious game. In S. A. Cerri & B. Clancey (Eds.), *Proceedings of the 11th international conference on intelligent tutoring systems (ITS 2012)* (pp. 162–167). Berlin, Germany: Springer.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education, 24,* 427–469.

OECD. (2013). *PISA 2015 collaborative problem solving framework*. (Paris, France: OECD) Retrieved from http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Olney, A., D'Mello, S. K., Person, N., Cade, W., Hays, P., Williams, C., & Graesser, A. C. (2012). Guru: A computer tutor that models expert human tutors. In S. Cerri, W. Clancey, G. Papadourakis & K. Panourgia (Eds.), *Proceedings of intelligent tutoring systems (ITS) 2012* (pp. 256–261). Berlin, Germany: Springer.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *LIWC2007: Linguistic inquiry and word count*. Austin, TX: LIWC.net.

Rosé, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., et al. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning, 3,* 237–271.

Rowe, J., Shores, L. R., Mott, B., & Lester, J. (2010). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education, 20,* 166–177.

Rus, V., D'Mello, S., Hu, X., & Graesser, A. C. (2013). Recent advances in intelligent systems with conversational dialogue. *AI Magazine, 34,* 42–54.

Rus, V., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2008). A study of textual entailment. *International Journal on Artificial Intelligence Tools, 17,* 659–685.

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: MIT Press.

Sottilare, R., Graesser, A., Hu, X., & Holden, H. (Eds.). (2013). *Design recommendations for intelligent tutoring systems: Learner modeling* (Vol. 1). Orlando, FL: Army Research Laboratory. Sottilare, Robert Graesser, Art Hu, Xiangen Holden, Heather.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. *Educational Psychologist, 46,* 197–221.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31,* 3–62.

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., & Weston, T. (2013). My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology, 105,* 1115–1125.

Zapata-Rivera, D., Jackson, G. T., & Katz, I. (2015). Authoring conversation-based assessment scenarios. In R. Sottilare, A. C. Graesser, X. Hu & K. Brawner (Eds.), *Design recommendations for intelligent tutoring systems* (Vol. 3, pp. 169–178). Orlando, FL: Army Research Laboratory.

Chapter 6
Assessment of Discussion in Learning Contexts

Carolyn Penstein Rosé, Iris Howley, Miaomiao Wen, Diyi Yang, and Oliver Ferschke

Abstract This chapter reports on our efforts to develop automated assessment of collaborative processes, in order to support effective participation in learning-relevant discussion. This chapter presents resources that can be offered to this assessment community by machine learning and computational linguistics. The goal is to raise awareness of opportunities for productive synergy between research communities. In particular, we present a three-part pipeline for expediting automated assessment of collaborative processes in discussion in order to trigger interventions, with pointers to sharable software and other opportunities for support. The pipeline begins with computational modeling of analytic categories, motivated by the learning sciences and linguistics. It also includes a data infrastructure for uniform representation of heterogeneous data sources that enables association between process and outcome variables. Finally, it includes supportive technologies that can be triggered through real-time, automated application of that analysis in order to achieve positive impact on outcomes.

Keywords Discourse analysis · Collaborative process analysis · Computational sociolinguistics · Text mining · Machine learning

C.P. Rosé (✉) · I. Howley · M. Wen · D. Yang · O. Ferschke
Language Technologies Institute and Human-Computer Interaction Institution,
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: cprose@cs.cmu.edu

I. Howley
e-mail: ihowley@cs.cmu.edu

M. Wen
e-mail: mwen@cs.cmu.edu

D. Yang
e-mail: diyiy@cs.cmu.edu

O. Ferschke
e-mail: ferschke@cs.cmu.edu

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_6

6.1 Introduction

Collaboration is a rich and multifaceted phenomenon, enacted through a complex interplay of multiple channels spanning a variety media. In this chapter, we focus on a single channel, collaboration through discussion. However, we also explore discussion across multiple settings where learning takes place, including classroom contexts, informal learning contexts, and Massive Open Online Courses (MOOCs). Discussion is the channel through which groups, teams, and communities work together to monitor and then maintain themselves, or to reconsider and then reset themselves, either through one-on-one interactions or public declarations. Discussion enables people to make their thinking public, which is a precondition for exchange of expertise and ideas.

Depending upon the context, the texture of the discussion may vary considerably. Thus, even within the single channel of discussion, analysis may take a wide variety of forms. Furthermore, insights and perspectives from multiple fields can be layered upon the rich but messy data trace left behind as collaboration unfolds. This chapter offers a perspective from machine learning and computational linguistics, not as a neatly packaged solution that can be used as a black box to solve a problem, but as an offer of collaboration with researchers of other fields that have their own equally valuable expertise to bring to the table. Researchers are invited to come together to discuss. Thus, the goal of this chapter is to raise awareness of the resources that can be offered from machine learning and computational linguistics to this assessment community. We position this chapter as just one contribution to what we hope will be an ongoing, collaborative conversation. Interested readers can continue the discussion by joining a community-building effort called DANCE: Discussion Affordances for Natural Collaborative Exchange[1], where we offer software resources, pointers into the literature, a monthly online interactive talk series, and opportunities to engage with the community in discussion through a Google group or Twitter.

The computational modeling efforts we report on in this chapter were developed to facilitate automated assessment of collaborative processes that are evident in discussion. That assessment is meant to provide a foundation for automated, dynamic support of collaborative discussion. In order to achieve that goal, we have adopted an input-process-output model, where our core research questions ask first what processes lead to which outcomes of interest, and next how those connections interact depending upon the preferred balance among multiple outcomes. Thus, we focus on measurement of both processes and outcomes as well as the connection between the two. In the supportive technologies we develop based on the insights from these modeling efforts, we can then dynamically trigger support from real-time analysis of process in a purposeful way. The goal is to achieve success in terms of outcomes we are concerned with.

[1] http://dance.cs.cmu.edu.

To this end, we present a three-part pipeline for expediting data analysis and student support. The pipeline begins with the data infrastructure for a uniform interface across heterogeneous data sources from a variety of discussion platforms where discussion for learning takes place. This infrastructure enables association between process and outcome variables, computational layering of analytic categories motivated by the learning sciences and linguistics, and supportive technologies that can be triggered through real-time application of that layered analysis. We present this three-part pipeline, and then conclude with caveats and directions for future work.

6.2 Operationalization of Processes and Outcomes

In our work, we focus first on acquisition of conceptual understanding and knowledge. We seek to understand how processes that occur through conversation contribute to success in these terms. We refer to this outcome as *learning*. As an important secondary outcome, we focus also on persistence-related variables, since students can only continue to increase their success at learning from their participation in a discussion if they continue to participate in it. We refer to this outcome as *commitment*.

The processes that most directly influence learning are cognitive. Thus, it would make sense to motivate analytic categories related to process in terms of theories of cognition. However, it is widely acknowledged that noncognitive factors such as social processes and dispositions influence the motivation to actively engage and persist in the cognitive processes most directly related to learning. Thus we must consider variables in other dimensions as well. Work on assessment of collaboration in the computer-supported collaborative learning community has therefore typically included dimensions for cognitive, social (or relational), and motivational processes (Strijbos, 2011).

In keeping with best practices in the field of computer-supported collaborative learning, the foundation for our computational work on assessment is a three-dimensional coding schema referred to as SouFLé (Howley, Mayfield, & Rosé, 2013); including cognitive, motivational, and social dimensions. It is designed to identify contributions that can be considered as signposts for sociocognitive conflict. The other two dimensions are meant to trace social positioning processes within conversation that move learners in and out of appropriate social proximity to one another for the purpose of facilitating engagement in the valued sociocognitive processes highlighted by the cognitive dimension.

SouFLé is offered as just an example including one operationalization of each of the three dimensions, where we have done computational modeling work both to automatically apply these codes as well as to measure how they relate to our outcomes of interest. SouFLé has been used in technology-supported analysis of chat data (Howley et al., 2012) as well as transcribed face-to-face discussion data (Ai, Kumar, Nguyen, Nagasunder, & Rosé, 2010). Its cognitive dimension has been applied to chat data (Joshi & Rosé, 2007), speech data (Gweon, Jain, Mc Donough, Raj, & Rosé, 2013) and transcribed discussion data from collaborative work (Gweon, Kane, & Rosé, 2011), and its motivational dimension has been applied to chat data (Howley, Mayfield, & Rosé, 2011) as well as transcribed face-to-face data (Mayfield, Laws, Wilson, & Rosé, 2014) as well. These studies serve as proofs of concept that operationalizations of the three dimensions can be computationalized successfully.

SouFLé has been used to analyze collaborative processes that have served as mediating variables explaining learning outcomes in collaborative learning settings (Howley et al., 2013; Howley, Mayfield, & Rosé, 2011). Analyses of related constructs have been used successfully to trigger automated forms of support for improving learning in collaborative settings (Ai et al., 2010; Kumar, Ai, Beuth, & Rosé, 2010). In the remainder of this section, we will explore each of these dimensions investigating their respective manifestations in different communication settings and how these process variables connect with outcomes.

6.2.1 Cognitive Process Variables

Howley et al. (2013) first introduced the SouFLé framework as a linguistic analysis approach for studying small groups. The intention was to define contribution-level codes in terms of basic language processes without reference to theoretical constructs that are specific to a particular theory of learning or collaboration. Instead, the goal was to ground the operationalizations in linguistics (Martin & Rose, 2003; Martin & White, 2005) and very broadly accepted learning-relevant constructs from the learning sciences (Berkowitz & Gibbs, 1979; Resnick, Asterhan, & Clark, 2015; Suthers, 2006; Teasley, 1997).

There is much evidence that something akin to the cognitive dimension in SouFLé is valuable as an assessment of the quality or effectiveness of episodes of collaborative learning. Across many different frameworks for characterizing discourse patterns associated with successful collaborative learning, the idea of eliciting articulation of reasoning and idea co-construction is a frequent central element (Chan, 2013; Chinn & Clark, 2013; van Alst, 2009), especially for its theoretical connection with cognitive conflict and learning (de Lisi & Golbeck, 1999). An example construct is that of transactivity (Berkowitz & Gibbs, 1979). Berkowitz and Gibbs defined a set of 18 different ways in which an articulation of reasoning can refer to or operate on the expressed reasoning of self or other. The expression of a transact reflects the examination of one's own mental model and possibly another's as well as the connections between them. Engaging in this process offers the opportunity for one to question one's own mental model. Thus, this key type of consensus-building behavior is theorized to play an important role in collaborative learning discourse. In Piaget's theory, these transactive contributions are most likely to occur within pairs or small groups where the participants are in an equal status relationship. Thus, we expect, and have found (Gweon et al., 2013), a connection between relational factors and the occurrence of transactive contributions. In our own prior work (Joshi & Rosé, 2007) as well as that of others (Azimitia & Montgomery, 1993; Weinberger & Fischer, 2006), we see evidence that prevalence of transactivity in collaborative discussions correlates with learning. Beyond a means for triggering cognitive conflict, transactive conversational contributions are viewed within this community as important steps in a knowledge-building or consensus-building process (Weinberger & Fischer, 2006). In making connections between newly articulated ideas and material contributed earlier in a conversation, ideas build upon one another, and differing understandings are elaborated, integrated, and possibly transformed. Prevalence of transactivity has also been demonstrated to correlate with successful transfer of expertise in loosely coupled work settings (Gweon et al., 2011).

In this spirit, the cognitive dimension of SouFLé is an operationalization of transactivity (Berkowitz & Gibbs, 1979; Weinberger & Fischer, 2006). It is distinct from the other two SouFLé dimensions in that its definition is not strictly linguistic. However, the values underlying the construct of transactivity (Berkowitz & Gibbs, 1979) are not really specific to a single theory of learning. The simple idea behind the concept of transactivity is a value placed on making reasoning explicit and elaborating expressed reasoning by building on or evaluating instances of expressed reasoning that came earlier in the discussion. The basic premise was that a reasoning statement should reflect the process of drawing an inference or conclusion through the use of reason. Statements that display reasoning can be coded as either externalizations, which represent a new direction in the conversation, not building on prior contributions, or transactive contributions, which operate on or build on prior contributions. In our distinction between externalizations and transactive contributions, we have attempted to take an intuitive approach by determining whether a contribution refers linguistically in some way to a prior statement, such as through the use of a pronoun or deictic expression. A recent analysis in a MOOC context demonstrates that prevalence of similar constructive explanation behavior predicts both learning (Wang, Wen, & Rosé, 2016; Wang, Yang, Wen, Koedinger, & Rosé, 2015) and commitment (Wen, Yang, & Rosé, 2014a).

6.2.2 Social Process Variables

The relational dimension in SouFlé is meant to capture the level of openness to the ideas of others that is communicated in a student's framing of assertions. Whereas in the cognitive dimension we adopted an approach in which we read into the text in order to identify expressions of reasoning and transactivity, in the relational dimension, we base our work on the earlier systemic functional linguistic (SFL) work of Martin and White (2005), whose theoretical approach explicitly mandates not going beyond the evidence that is explicit in a text. The important distinction in our application of Martin and White's heteroglossia framework is the distinction between a monoglossic assertion, which is framed as though it leaves no room for questioning, in contrast to those framed in a heteroglossic manner, where the assumed perspective of others is explicitly acknowledged within the framing. There are two types of contributions we code as heteroglossic, one type that shows openness to other perspectives, which we refer to as *heteroglossic expand*, and another that explicitly expresses a rejection of some other perspective, which we refer to as *heteroglossic contract*. In our work, in both correlational and experimental studies, we have found that concentration of heteroglossic expand statements within an interaction significantly predicts the articulation of reasoning (Dyke, Howley, Adamson, Kumar, & Rosé, 2013; Kumar, Beuth, & Rosé, 2011). This empirical evidence supports the importance of including a dimension like this within the framework. This evidence supports the claim that this social construct supports an important cognitive construct.

In our work, other social variables have also been demonstrated to make significant predictions about commitment in MOOCs. We have applied machine learning in a MOOC context to detect emergent subcommunities in MOOC discussion forums (Rosé et al., 2014), student attitudes towards course affordances and tools (Wen, Yang, & Rosé, 2014b), satisfaction with help received (Yang, Wen, & Rosé, 2014b), and participation in discussion threads, interests, and relationship formation (Yang, Wen, & Rosé, 2014a). All of these make significant predictions about commitment based on survival analyses.

6.2.3 Motivational Process Variables

The motivational dimension in SouFLé is meant to capture conversational behavior that reflects the self-efficacy of students related to their ability to participate meaningfully in the collaborative learning interaction (Howley et al., 2011). In our prior work we have seen correlations between self-report measures of collective self-efficacy from collaborative groups and measures of authoritativeness of stance derived from our coding of this dimension. We have also found a mediating effect with measures of learning (Howley et al., 2012). In short, on this dimension we consider that an authoritative presentation of knowledge is one that is presented without seeking external validation for that knowledge. This dimension, which we have referred to as the *authoritativeness framework*, is rooted in Martin's negotiation framework (Martin & Rosé, 2003), from the systemic functional linguistics community. This framework highlights the moves that are made in a dialogue as they reflect the authoritativeness with which those moves were made, and gives structure to exchanges back and forth between participants. Application of this dimension has been successfully automated in chat (Howley et al., 2012), transcribed doctor-patient interactions (Mayfield et al., 2014), and transcribed collaborative discussions (Mayfield & Rosé, 2011).

6.3 DiscourseDB: A Data Infrastructure for Bringing Multiple Data Streams Together

The foundation of computational analytic work is representation of data. Much of our published work in assessment of collaboration in discussion has been focused on either chat data or transcribed face-to-face discussion. These can both be represented in a simple, uniform, flat sequence of text segments, each contributed by one speaker. However, when expanding to learning in MOOCs or learning in other online contexts such as open-source communities, the form that the discussions may take becomes more diverse as they are embedded in a variety of platforms. They may even occur simultaneously through multiple separate streams. To that end we offer a publically available data infrastructure we call DiscourseDB,[2] which enables translation of data from multiple streams into a common, integrated representation.

As a concrete example, consider connectivist Massive Open Online Courses (cMOOCs) that include environments like the competency-based learning platform ProSolo (Jo, Tomar, Ferschke, Rosé, and Gaesevic in press). In these environments, data are rich and heterogeneous. In ProSolo, for example, student behaviors formally within the environment include follower-followee relations, posting wall notes including updates and goal notes, and commenting on notes. Students also engage in threaded discussions, blog and comment on blog posts, and tweet. These behaviors occur within accounts in other linked online community spaces. In a proof of concept using data from the edX Data, Analytics, and Learning course,[3] we have transformed data from wall post comments, blogs and blog comments, and Twitter into DiscourseDB and applied probabilistic graphical modeling techniques to identify typical student learning trajectories that could be supported through social recommendation (Jo et al., in press).

The goal of DiscourseDB is to facilitate analysis of discussion data across multiple platforms. Specifically, we are developing DiscourseDB, which is a database infrastructure that is capable of accommodating threaded discussion, chat, blogs with comments, e-mail and personal messaging, Twitter and other microblogs, as well as wikis (including their talk pages). It accommodates data that can be scraped from the source platforms or exported in the form of data dumps. These platforms are different in terms of what is explicit, what is implicit (but retrievable), and what is implicit (and not retrievable). And our representation enables us to store discourse data in a common format in ways that respect these differences.

In the DiscourseDB representation, discourse is broken down into its macro structure and its micro structure. On the content level, the macro structure is represented in a relational database as an entity-relation model of connected discourse contributions organized in generic, nested discourse containers. These containers

[2] https://discoursedb.github.io/.

[3] https://www.edx.org/course/data-analytics-learning-utarlingtonx-link5-10x.

capture the organizational structures of the different source platforms, such as forums, subforums, threads, chats, or discussion pages. Relationships between contributions can be arbitrarily typed and thus make it possible to represent both explicit and implicit properties under the same paradigm.

The user level of the macro structure represents both the actively and passively involved individuals, that is, the authors or revisers of contributions and their audiences. Each instantiation of a contribution is associated with its author, thereby resulting in a set of users involved in the creation and revision of a contribution over time. Users can be organized in groups of arbitrary types. These groups can represent teams in team-based collaboration platforms, but also resemble formal role-based aggregations such as groups with different access rights or status on the source platform. The micro structure captures the internal organization of individual contributions using the Unstructured Information Management Architecture (UIMA) (Ferrucci & Lally, 2004).

6.4 Connecting Processes with Outcomes

Earlier we discussed coding schemas relating to three dimensions of collaboration, and how sums and proportions of the codes within those schemas make predictions about learning and commitment. In our recent work, we have moved beyond consideration of concentration of individual codes to thinking in terms of roles in collaboration that are defined based on characteristic distributions of codes, or behavior profiles. The work of discussion occurs not just as the sum of the effects of each individual contribution or single type of behavior. It is rather a result of enactment of roles working together over periods of time. In this spirit, in this section we describe a novel approach to identifying behavior profiles that make predictions about outcomes. We describe how the resulting models of role taking associated with outcomes can be used to trigger interventions that support effective role taking in collaboration. We refer to this modeling framework as the role identification model (RIM) because we conceptualize the distributions of behaviors identified as valuable for achieving outcomes as descriptions of roles that participants take on within interaction.

In two successful proofs of concept, our initial work with RIM was applied to the problem of predicting group grades in team-based MOOCs (Yang, Wen, & Rosé, 2015) and behavior profiles within discussion pages that predict page quality in Wikipedia (Ferschke, Yang, & Rosé, 2015). Our RIM model aims to maximize the predicted quality scores of teamwork among a selected set of key participants. This modeling framework links a representation of interaction processes with outcomes and thus provides the foundation for the proposed modeling work, in which we explore extensions to this framework. In this work, each person's behavior representation is a vector, where each feature is a count or proportion of some code, such as the cognitive, social, and motivational categories described above.

Here we first introduce the basic notation and then present a qualitative description of an iterative process for identification of role-based behavior profiles. Suppose we have C teams (or some other social unit) in which participants collaborate to achieve an outcome. The number of participants in the j-th team is denoted as Nj, (1 ≤ j ≤ Nj). There are K roles across C teams that we want to identify, where 1 ≤ K ≤ N_j; for all j in [1,C]. That is, the number of roles is smaller than or equal to the number of participants in a team, which means that each role should have one participant assigned to it, but not every user needs to be assigned to a role. Each role is associated with a weight vector W_k in RD to be learned, 1 ≤ k ≤ K and D is the number of dimensions. Each participant i in a team j is associated with a behavior vector B_j,i in RD. The measurement of teamwork quality is denoted as Q_j for team j, and the predicted Q_j is determined by the inner product of the behavior vectors of participants who are assigned to different roles and the corresponding weight vectors. The goal of the modeling process is to find a proper teamwork role assignment that positively contributes to the teamwork outcome (i.e., improvement of article quality or a high grade on a group project) as much as possible.

The role identification process is iterative and involves two stages. The first stage uses a regression model to adjust the weight vectors in order to predict the teamwork quality, given a fixed role assignment that assumes participants are well matched to roles. In the second stage, we iterate over possible assignments to find a match between participants and roles that maximizes our objective measure. In order to avoid the complexity of a brute force enumeration, we create a weighted bipartite graph and apply a maximum weighted matching algorithm (Ahuja, Magnanti, & Orlin, 1993) to find the best match. For each team, a separate graph is created. We alternate between the two stages until both role assignment and teamwork quality prediction converge.

6.5 Triggering Support for Collaboration from Automated Assessment

The end goal of our modeling work is to produce design recommendations for interventions that can be used to increase the level of supportiveness within online learning communities. We see the technology we aim to design as an augmentation of human effort, where the technology provides feedback and guidance to individuals and groups in order to enable social units to self-regulate, develop, and improve. The identification of sets of behavior profiles that together predict outcomes forms the foundation for future interventions that have the potential to guide participants to opportunities where they can productively contribute towards positive outcomes based on their observed past behavior. We envision that many of these interventions will be implemented in the form of social recommendation approaches (Jo et al., in press; Yang & Rosé, 2014c)

Our prior work has already produced suggestive evidence that a social recommendation approach, triggered based on automated analyses of community engagement, can lead to positive impact in online learning contexts. For example, Yang & Rosé, (2014c) employed a feature-aware matrix factorization approach that was used to identify behavior profiles that predict goodness of fit between a discussion opportunity and a participant's past observed discussion behavior. In that work, latent behavior profiles of individual users were used to rank opportunities for participation in terms of goodness of fit between the needed profile and the observed profile of each user. A constraint satisfaction approach was used for load balancing so that available human resources were effectively dispatched throughout the community. That algorithm was used in a successful help-seeking support intervention referred to as the Quick Helper, which was deployed in the edX Data, Analytics, and Learning MOOC (DALMOOC; Howley, Tomar, Yang, Ferschke, & Rosé, 2015). The goal of Quick Helper was to match help seekers with help providers.

In a post hoc analysis of DALMOOC (Jo et al. in press), the three-part pipeline described in this chapter was used as a foundation for a social recommendation approach to improve the follower-followee network of students in the course in order to increase the extent to which students had access to role models exhibiting effective goal-directed behavior. Suggestive evidence from a corpus-based evaluation was provided.

Going forward, we propose to design social recommendation interventions of a similar nature, but using extensions of RIM rather than the feature-aware matrix factorization approach used in our prior work. The RIM framework is more versatile for identifying behavior profiles that form the basis for recommendation for our proposed purposes, because of the natural way it enables us to drive the process of identification of key behavior profiles from desired outcomes and the way we propose to extend it in order to account for the different ways support might be targeted.

6.6 Caveats and Current Directions

In this chapter we have motivated and described a three-part pipeline for expediting automated assessment of collaborative processes in discussion, in service of triggering interventions. We do not by any means claim that our computational modeling research is done. Instead we present this chapter as an overview of our progress so far and invite collaboration with researchers in the assessment community who would like to join forces in this effort.

We acknowledge significant opportunities for improvement in our current approaches. For example, in our current work with RIM, we construct behavior vectors where each type of behavior is a feature, and the value of that feature is the number or proportion of such behaviors contributed by the participant, as we did in the Wikipedia work (Ferschke et al., 2015) as well as the team-based MOOC work

(Yang et al., 2015). This way of characterizing a discussion process can be thought of as treating conversation as a container for collecting unordered sets of behaviors and is sometimes referred to as "coding and counting." It is a typical way of characterizing a discussion process in behavioral research, but it has been criticized as having important limitations. In particular, it misses the way communicative behaviors within groups may be targeted at the whole group, or individuals within the group, or even individuals outside of the group, although these distinctions may have important consequences for the functioning of the group. One important limitation of coding and counting approaches is that they collapse a participant's contributions over time and thus render the participant's change over time invisible to the model. In future work, we plan to overcome this limitation by instead conceptualizing interaction as a developmental process. What this means is that we will relax the assumption that there is a persistent one-to-one correspondence between participants and roles over time within an interaction. Instead, we will employ a mixed-membership approach that allows roles to be played by different combinations of people over the course of the interaction and recognizes that the same person may at different times contribute to the execution of different roles.

Another limitation of the coding and counting approach is that each behavior is treated separately, and thus contingencies between contributions as they occur over time are invisible to the model. Modeling these connections is an important precursor to being able to take account of the intended audience of a contribution. Prior work in conversation analysis points to ways that even basic conversational functions like object reference are accomplished jointly by participants (Clark & Bresnan, 1991; Clark & Schaefer, 1989). Another important way that behavior is contingent in collaborative work is that leadership behavior only contributes to group outcomes if team members respond to the leader's behaviors. For example, a leader may assign tasks to people who have the requisite skills to carry out the work most effectively, but if those people choose not to take up the assignments, the work will not get done successfully. Thus, in future work, rather than treating behaviors and roles as contributing independently to outcomes, we will seek to represent behaviors in ways that enable leveraging contingencies, and extend the modeling framework to take advantage of this representation.

A final limitation of the coding and counting approach is that each contribution is treated as though it has a persistent status; however, that may not be true where contributions can be coauthored or edited, as is the case both in Wikipedia discussion pages as well as notes in environments such as the Knowledge Forum (Scardamalia & Bereiter, 1993, 2006). Thus, in future work, we will extend the developmental process approach still further by relaxing the assumption that contributions maintain the same status in terms of what behaviors are associated with them over time.

Beyond the extensions to our modeling work that we propose to contribute going forward, the more important next step is to engage transactively with other communities of assessment researchers. We hope this chapter will kick off that exchange.

Acknowledgements This work was funded in part by NSF grants ACI-1443068 and OMA-0836012 and funding from Google.

References

Ahuja, R. K., Magnanti, T. L., & Orlin, J. B. (1993). *Network flows: Theory, algorithms, and applications*. Englewood Cliffs, NJ: Prentice Hall.

Ai, H., Kumar, R., Nguyen, D., Nagasunder, A., & Rosé, C. P. (2010). Exploring the effectiveness of social capabilities and goal alignment in computer supported collaborative learning. *Lecture Notes in Computer Science, 6095,* 134–143.

Azmitia, M., & Montgomery, R. (1993). Friendship, transactive dialogues, and the development of scientific reasoning. *Social Development, 2*(3), 202–221.

Berkowitz, M., & Gibbs, J. (1979). *A preliminary manual for coding transactive features of dyadic discussion*. Unpublished manuscript, Marquette University, Milwaukee, WI, 2(1), 6–1.

Chan, C. K. K. (2013). Collaborative knowledge building: Towards a knowledge creation perspective. In C. E. Hmelo-Silver, C. A. Chinn, C. K. K. Chan & A. M. O'Donnell (Eds.), *International handbook of collaborative learning* (pp. 437–461). New York, NY: Taylor and Francis.

Chinn, C. A., & Clark, D. B. (2013). Learning through collaborative argumentation. In C. E. Hmelo-Silver, C. A. Chinn, C. K. K. Chan & A. M. O'Donnell (Eds.), *International handbook of collaborative learning* (pp. 437–461). New York, NY: Taylor and Francis.

Clark, H., & Bresnan, J. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine & S. D. Teasley (Eds.), *Perspectives on socially shared cognition*, (pp. 127–149). Washington, DC: American Psychological Association.

Clark, H., & Schaefer, E. (1989). Contributing to discourse. *Cognitive Science, 13*(2), 259–294.

de Lisi, R., & Golbeck, S. L. (1999). Implications of the Piagetian theory for peer learning. In A. M. O'Donnell & A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 3–37). Mahwah, NJ: Lawrence Erlbaum Associates.

Dyke, G., Howley, I., Adamson, D., Kumar, R., & Rosé, C. P. (2013). Towards academically productive talk supported by conversational agents. In D. D. Suthers, K. Lund, C. P. Rosé, C. Teplovs & N. Law (Eds.), *Productive multivocality in the analysis of group interactions* (pp. 459-476). New York, NY: Springer.

Ferschke, O., Yang, D., & Rosé, C. P. (2015). A lightly supervised approach to role identification in Wikipedia talk page discussions. In *International AAAI conference on web and social media.* Retrieved March 29, 2016 from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10634

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering, 10*(3–4), 327–348.

Gweon, G., Jain, M., Mc Donough, J., Raj, B., & Rosé, C. P. (2013). Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. *International Journal of Computer Supported Collaborative Learning, 8*(2), 245–265.

Gweon, G., Kane, A., & Rosé, C. P. (2011, July). *Facilitating knowledge transfer between groups through idea co-construction processes*. Paper presented at the annual meeting of the Interdisciplinary Network for Group Research (INGRoup), Minneapolis, MN.

Howley, I., Mayfield, E., & Rosé, C. P. (2011). Missing something? Authority in collaborative learning. *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL 2011 Conference Proceedings—Long Papers, 9th International Computer-Supported Collaborative Learning Conference* (Vol. 1, pp. 366–373). Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84858400613&partnerID=tZOtx3y1

Howley, I., Adamson, D., Dyke, G., Mayfield, E., Beuth, J., & Rosé, C. P. (2012). Group composition and intelligent dialogue tutors for impacting students' self-efficacy. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7315,* 551–556.

Howley, I., Tomar, G., Yang, D., Ferschke, O., & Rosé, C. P. (2015). Alleviating the negative effect of up and downvoting on help seeking in MOOC discussion forums. In *Proceedings of the 17th international conference on artificial intelligence in education (AIED 2015),* IOS Press.

Howley, I., Mayfield, E., & Rosé, C. P. (2013). Linguistic analysis methods for studying small groups. In C. Hmelo-Silver, A. O'Donnell, C. Chan & C Chin (Eds.) *International handbook of collaborative learning*. London: Taylor and Francis, Inc.

Jo, Y., Tomar, G., Ferschke, O., Rosé, C. P., & Gaesevic, D. (in press). Pipeline for expediting learning analytics and student support from data in social learning. *Proceedings of the 6th international learning, analytics, and knowledge conference (LAK16)* (poster).

Joshi, M., & Rosé, C. P. (2007, October). Using transactivity in conversation summarization in educational dialog. *Proceedings of the ISCA special interest group on speech and language technology in education workshop (SLaTE),* Farmington, PA. Retrieved from http://www.isca-speech.org/archive_open/archive_papers/slate_2007/sle7_053.pdf

Kumar, R., Beuth, J., & Rosé, C. P. (2011). Conversational strategies that support idea generation productivity in groups. *Connecting computer-supported collaborative learning to policy and practice: CSCL 2011 Conference Proceedings—Long Papers, 9th International Computer-Supported Collaborative Learning Conference* (Vol. 1. pp 398–405).

Martin, J. R., & Rose, D. (2003). *Working with discourse: Meaning beyond the clause*. New York, NY: Continuum.

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. New York: Palgrave/Macmillan.

Mayfield, E., Laws, B., Wilson, I., & Rosé, C. P. (2014). Automating annotation of information flow for analysis of clinical conversation. *Journal of the American Medical Informatics Association, 21*(1), 122–128.

Mayfield, E., & Rosé, C. P. (2011). Recognizing authority in dialogue with an integer linear programming constrained model. In *Proceedings of the 49th annual meeting of the association for computational linguistics*, 1018–1026. Retrieved from http://www.aclweb.org/anthology-new/P/P11/P11-1102.pdf

Resnick, L., Asterhan, C., & Clarke, S. (2015). *Socializing intelligence through academic talk and dialogue*. Washington, DC: American Educational Research Association.

Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P., & Sherer, J. (2014). Social factors that contribute to attrition in MOOCs. In *Proceedings of the first ACM conference on learning @ Scale*. - L@S '*14*, 197–198. Retrieved from http://dl.acm.org/citation.cfm?id=2556325.2567879

Scardamalia, M., & Bereiter, C. (1993). Technologies for knowledge-building discourse. *Communications of the ACM, 36*(5), 37–41.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 97–118). New York, NY: Cambridge University Press.

Strijbos, J. W. (2011). Assessment of (computer-supported) collaborative learning. *IEEE Transactions on Learning Technologies, 4*(1), 59–73.

Suthers, D. (2006). Technology affordances for inter-subjective meaning making: A research agenda for CSCL. *International Journal of Computer Supported Collaborative Learning, 1*(3), 315–337.

Teasley, S. D. (1997). Talking about reasoning: How important is the peer in peer collaborations? In L. B. Resnick, R. Saljo, C. Pontecorvo & B. Burge (Eds.), *Discourse, tools, and reasoning: Situated cognition and technologically supported environments* (pp. 361–384). Heidelberg, Germany: Springer-Verlag.

van Aalst, J. (2009). Distinguishing between knowledge sharing, knowledge creating, and knowledge construction discourses. *International Journal of Computer Supported Collaborative Learning, 4*(3), 259–288.

Wang, X., Wen, M., & Rosé, C. P. (2016). Towards triggering higher-order thinking behaviors in MOOCs. In *Proceedings of the 6th international learning, analytics, and knowledge conference (LAK16).*

Wang, X., Yang, D., Wen, M., Koedinger, K. R., & Rosé, C. P. (2015). Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains. In *Proceedings of the 8th international educational data mining conference (EDM15).*

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46,* 71–95.

Wen, M., Yang, D., & Rosé, D. (2014a). Linguistic reflections of student engagement in massive open online courses. In *Proceedings of the international conference on weblogs and social media.* Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8057/8153

Wen, M., Yang, D., & Rosé, C. P. (2014b). Sentiment analysis in MOOC discussion forums: What does it tell us? In *Proceedings of the 7th international educational data mining conference (EDM14).*

Yang, D., Wen, M., & Rose, C. P. (2014a). Towards identifying the resolvability of threads in MOOCs. In Association for computational linguistics (Ed.), *Proceedings of the EMNLP workshop on modeling large scale social interaction in massively open online courses* (pp. 21–31). Doha, Qatar: Association for Computational Linguistics.

Yang, D., Wen, M., & Rosé, C. P. (2014b). Peer influence on attrition in massively open online courses. In *Proceedings of the 7th international educational data mining conference (EDM14).*

Yang, D. & Rosé, C. P. (2014c). Constrained question recommendation in MOOCs via submodality, *Proceedings of the 2014 ACM international conference on information and knowledge management,* pp. 1987–1990.

Yang, D., Wen, M., & Rosé, C. P. (2015). Weakly supervised role identification in teamwork interactions. In *Proceedings of the 53rd annual meeting of the association for computational linguistics.*

Chapter 7
Collaborative Problem Solving Measures in the Programme for International Student Assessment (PISA)

Qiwei He, Matthias von Davier, Samuel Greiff, Eric W. Steinhauer, and Paul B. Borysewicz

Abstract Collaborative problem solving (CPS) is a critical and necessary skill in educational settings and the workforce. The assessment of CPS in the Programme for International Student Assessment (PISA) 2015 focuses on the cognitive and social skills related to problem solving in collaborative scenarios: establishing and maintaining shared understanding, taking appropriate actions to solve problems, and establishing and maintaining group organization. This chapter draws on measures of the CPS domain in PISA 2015 to address the development and implications of CPS items, challenges, and solutions related to item design, as well as computational models for CPS data analysis in large-scale assessments. Measuring CPS skills is not only a challenge compared to measuring individual skills but also an opportunity to make the cognitive processes in teamwork observable. An example of a released CPS unit in PISA 2015 will be used for the purpose of illustration. This study also discusses future perspectives in CPS analysis using multidimensional scaling, in combination with process data from log files, to track the process of students' learning and collaborative activities.

This work was conducted while Matthias von Davier was employed with Educational Testing Service.

Q. He (✉) · E.W. Steinhauer · P.B. Borysewicz
Educational Testing Service, Princeton, NJ, USA
e-mail: qhe@ets.org

E.W. Steinhauer
e-mail: esteinhauer@ets.org

P.B. Borysewicz
e-mail: pborysewicz@ets.org

S. Greiff
Computer-Based Assessment Research Group, University of Luxembourg,
Luxembourg City, Luxembourg
e-mail: samuel.greiff@uni.lu

M. von Davier
National Board of Medical Examiners (NBME), Philadelphia, USA
e-mail: mvondavier@gmail.com

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_7

Keywords Collaborative problem solving · Computer-based assessment · Large-scale assessment · Test development · Item response theory · PISA

7.1 Introduction

In the 21st century, the types of skills needed to succeed have undergone a rapid and substantial change. Factual knowledge that was at the core of many professions a mere half century ago is virtually immediately available in the 21st century with the advent of the Internet. Also, noncognitive skills that intersect with cognitive ones now involve mastering new challenges and require cooperative efforts among a group of individuals. Such skills are increasingly needed to meet the demands of the 21st century, whether in education, at the workplace, or elsewhere in life (Griffin et al., 2012; Greiff et al., 2014). For instance, Autor, Levy, and Murnane (2003) highlighted that manual and routine cognitive tasks have been declining in importance and frequency across professions over the last decades (Cascio, 1995; Goos, Manning, & Salomons, 2009), while at the same time we have increasingly faced challenges that have not been encountered before and that require cooperation to solve efficiently.

Skillfully dealing with new problems in diverse settings and contexts, as part of a team instead of individually, is at the core of the concept of collaborative problem solving (CPS). CPS reflects a set of skills that combines cognitive and social aspects that are relevant for successful problem solving across domains regardless of the specific contextual setting. Importantly, one of the most acknowledged educational large-scale assessments, the Programme for International Student Assessment (PISA), which is organized by the Organisation for Economic Co-operation and Development (OECD), complemented its assessment portfolio with a fully computer-based assessment of CPS in the 2015 cycle (OECD, 2013). The triennial PISA study measures proficiency levels of 15-year-old students in over 70 countries, including OECD members as well as nonmember countries (known as partner countries), in the core domains of mathematics, science, and reading. Previous PISA cycles already included measures of skills that intersect with the core domains, specifically individual problem solving in PISA 2003 (paper-and-pencil-based) and 2012 (computer-based), acknowledging these skills' increasing relevance. A bold move was made in PISA 2015 as CPS was included for the first time, explicitly incorporating both social and cognitive aspects in the assessment. Such innovation introduces a new viewpoint to understanding students' performance proficiency that goes beyond the borders of domain-specific competencies and mere cognitive ability constructs such as reasoning and working memory (Greiff et al., 2014).

7.1.1 CPS Framework in PISA

Selected by the OECD as an innovative domain to be assessed in PISA 2015, CPS is defined in the draft framework as "the capacity of an individual to effectively

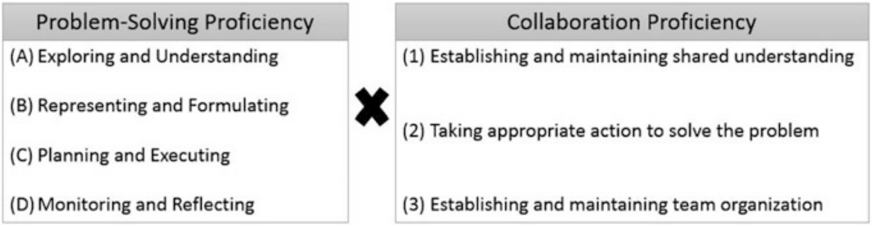

Problem-Solving Proficiency
(A) Exploring and Understanding
(B) Representing and Formulating
(C) Planning and Executing
(D) Monitoring and Reflecting
Collaboration Proficiency
(1) Establishing and maintaining shared understanding
(2) Taking appropriate action to solve the problem
(3) Establishing and maintaining team organization

Fig. 7.1 Two core proficiencies in the PISA CPS framework (OECD, 2013)

engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution" (OECD, 2013, p. 6). It was designed specifically for the computer-based assessment (CBA) mode. Two core domains are involved: problem-solving proficiency (a mainly cognitive domain including four dimensions) and collaboration proficiency (a mainly social domain including three dimensions), thus tapping simultaneously into both (see Fig. 7.1). In combination, these two domains reflect students' CPS proficiency.

The left panel in Fig. 7.1 displays four cognitive dimensions in individual problem solving: exploring and understanding, representing and formulating, planning and executing, and monitoring and reflecting (OECD, 2013). These dimensions are consistent with the PISA 2012 problem-solving framework. A similar set of dimensions was also identified in the Programme for the International Assessment of Adult Competencies (PIAAC) problem solving in technology-rich environments framework, which focuses more on processes related to the acquisition, use, and production of information in computerized environments (OECD, 2009). The CPS framework in PISA was developed based on the previous assessments of individual problem solving with an additional integration of collaborative elements (OECD, 2013).

In the problem-solving domain, the first cognitive dimension (exploring and understanding) involves understanding the situation where a problem is encountered by interpreting initial information about it and any information uncovered during exploration and interactions with the problem. The second dimension (representing and formulating) involves selecting, organizing, and integrating relevant information with prior knowledge. In this process, information is initially presented by graphs, tables, symbols, and words. Hypotheses may be formulated based on identification of problem factors and evaluation of critical information. The third dimension (planning and executing) includes planning, which consists of clarifying the goal of the problem, setting any subgoals, and developing a plan to reach the goal. Executing that plan is also part of this process. The final dimension (monitoring and reflecting) involves monitoring and reflecting on one's actions and is related to changing actions and strategies throughout the problem-solving process. These four problem-solving dimensions are the foundation for developing an assessment framework for individual's problem-solving skills and provide the possibility of structuring with a joint assessment dimension in collaborative process.

As for the social aspects, the PISA 2015 framework incorporated them into the four problem-solving dimensions by focusing on three major dimensions of collaboration, which are shown in the right panel in Fig. 7.1. In accordance with the guidelines of CPS drafted by the OECD (2013), in the first dimension under collaboration (establishing and maintaining shared understanding), students are required to show their abilities to identify mutual knowledge, understand perspectives of peers (other agents in the collaboration), and form a common understanding about the problem (OECD, 2013). Students also need to use effective means of communication, for instance, responding to requests, sending information to peers about the process for joint tasks, sharing knowledge, confirming what has been understood by each other, taking actions to clarify misunderstandings, and so on. These skills focus on students' self-awareness and awareness of others' proficiencies in performing a task, that is, recognizing their own and their peers' strengths and weaknesses in relationship to the task (Cannon-Bowers & Salas, 2001; Dillenbourg, 1999; Dillenbourg & Traum, 2006; Fiore & Schooler, 2004). The second dimension (taking appropriate action to solve the problem) emphasizes a joint effort that takes group members' specific skill profiles and external constraints into account and monitors the process to achieve the group goal. Communication actions such as explaining, debating, arguing, and negotiating are involved in order to transfer information and find more optimal solutions (OECD, 2013). The third dimension (establishing and maintaining team organization) focuses on students' capability of understanding different roles within the group based on their knowledge of each team member's skills, adjusting to changes, and monitoring the group organization (OECD, 2013).

Importantly, the PISA CPS framework assumes that in each collaborative problem-solving effort, one of the four problem-solving dimensions and one of the three collaboration dimensions are central for success or failure, whereas the other dimensions play only a minor role. Therefore, the problem-solving and collaboration dimensions are combined into a set of 12 (4 × 3) detailed skills (see Fig. 7.1). For instance, the combination of A1 (i.e., (A) exploring and understanding in problem-solving proficiency combined with (1) establishing and maintaining shared understanding in collaboration proficiency) is used to assess students' skills in discovering perspectives and abilities of team members.

The PISA CPS units were developed in a way that ensured that the 12 CPS skills were all measured across different tasks (OECD, 2013). Each skill, representing the intersection of one of the three collaboration dimensions and one of the four problem-solving dimensions, was based upon the rich body of research that exists in fields such as problem solving, cognitive psychology, collaborative learning, and so forth, even though the terms used in the PISA CPS framework might deviate from those in the scientific literature.[1]

[1] For instance, "planning and executing" is usually referred to as *knowledge application* in complex problem-solving research; see Wüstenberg, Greiff, and Funke (2012).

7.1.2 CPS Test Development in PISA

This CPS framework consequently served as the basis for test development of CPS units[2] designed for the PISA target population. In fact, it was the aim of the test developing process to comprehensively cover the set of 12 framework skills in units that were appropriate for 15-year-olds attending school. Obviously, this required tapping into both the cognitive and the social dimensions of CPS in order to broaden the view on students' proficiency levels. Thus, the endeavor of measuring CPS in an international large-scale assessment such as PISA was a challenge; few comparable efforts have been conducted before. The Assessment and Teaching of 21st Century Skills (ATC21S) initiative (Griffin et al., 2012) assessed collaboration in problem-solving environments in a couple of countries, but the international coverage of CPS in PISA 2015 was unprecedented. Moreover, assessments such as PISA operate under a number of constraints that require potentially impactful choices with regard to the nature of the assessment. To this end, a priority in test development for PISA 2015 was placed on standardization in order to obtain comparable scores across students from a wide range of countries. Standardization had significant implications for the way students interacted with the other team members in the CPS units.

In order to ensure that students experienced communication patterns and collaborative behaviors with peers that were comparable, the PISA CPS items were designed as students interacted with computer-simulated agents and communicated with the agents through predefined chat messages. Through this process, all students encountered largely the same stimuli and had the same opportunities to react as they worked to solve the problem-solving situation. It was unlike the ATC21S assessment mentioned earlier, which put stronger emphasis on interaction and collaboration between two humans in a free-chat environment, employing a human-human approach instead of the human-agent approach chosen in PISA. On the one hand, the ATC21S approach increases the resemblance of the assessment to real-world interactions, but on the other, it limits standardization of the assessment and makes it difficult to score individual performance. In this juxtaposition of internal validity (e.g., standardization through human-agent interaction) and external validity (e.g., real-world resemblance through human-human interaction), the PISA 2015 assessment set a high priority on the psychometric quality of the assessment, that is, emphasizing internal validity and comparability. However, in order to empirically verify sufficient external validity of the human-agent approach employed in PISA, the OECD is conducting studies that compare the human-human and the human-agent assessment approach as well.

Huge potential and significant challenges coexist in extending educational large-scale assessments toward social and noncognitive dimensions and making comparisons across countries in CPS skills. Although the inclusion of CPS

[2] In the PISA context, a task that, in turn, might be composed of several items is considered as one unit.

assessment in the PISA 2015 presented a leap of innovation toward an assessment of both cognitive and social skills, it is acknowledged that many questions still remain unanswered.

7.2 Construction of CPS Items for PISA 2015

7.2.1 Guidelines for CPS Item Development

In seeking to translate the CPS construct into a measurement instrument, the PISA Collaborative Problem Solving Expert Group determined that assessment units would present simulated problem-solving scenarios calling for the student to work with a small team trying to accomplish a common goal. A number of criteria were set in the item development process to optimize the assessment of CPS skills, of which the following were most important:

- Team members' information or roles were asymmetric, that is, different team members had different information, roles, or resources.
- Teams were presented with problems that allowed for more than one solution, meaning there was room to make different decisions in reaching a solution, leaving space for collaborative choices instead of the dictates of the requirements of a single solution.
- Information was provided dynamically. Rather than receiving all necessary problem-solving information at the outset, the student and the team received important information as the scenario unfolded.
- Team size was constrained at a maximum of five members, meaning the student worked with one to four teammates per unit. This made it easier for a student to keep track of the perspectives of teammates, which is a critical element in some CPS skills.

In addition, a set of guidelines in task and scenario selection was recommended by the expert group to ensure accessibility for the full range of PISA test takers. For instance, each scenario set up team members as peers instead of having hierarchical authorities. Also, the task selection favored practical problems over academic work to reduce variances in performance associated with specific academic content. Lastly, it was important to avoid interactions where nuance of tone or word choice might create misunderstanding when translating between different languages.

7.2.2 CPS Item Design

In the PISA 2015 CPS assessment, a closed answering format was used. Students had to navigate through a collaborative environment (i.e., units) and master subtasks in each environment (i.e., items) by choosing the best alternative out of sets of predefined answers and communication options. We will use a released CPS unit (The Visit) in this chapter to illustrate how a CPS unit looks and how such items are developed. This unit, including three parts and 44 measurable items, was completed by students during the PISA 2015 field trial in an average of 17 minutes. The premise for this unit was that a group of international students was coming to visit a school. The respondent had to collaborate with three agent teammates to plan the visit, assign visitors to guides, and respond to an unexpected problem.

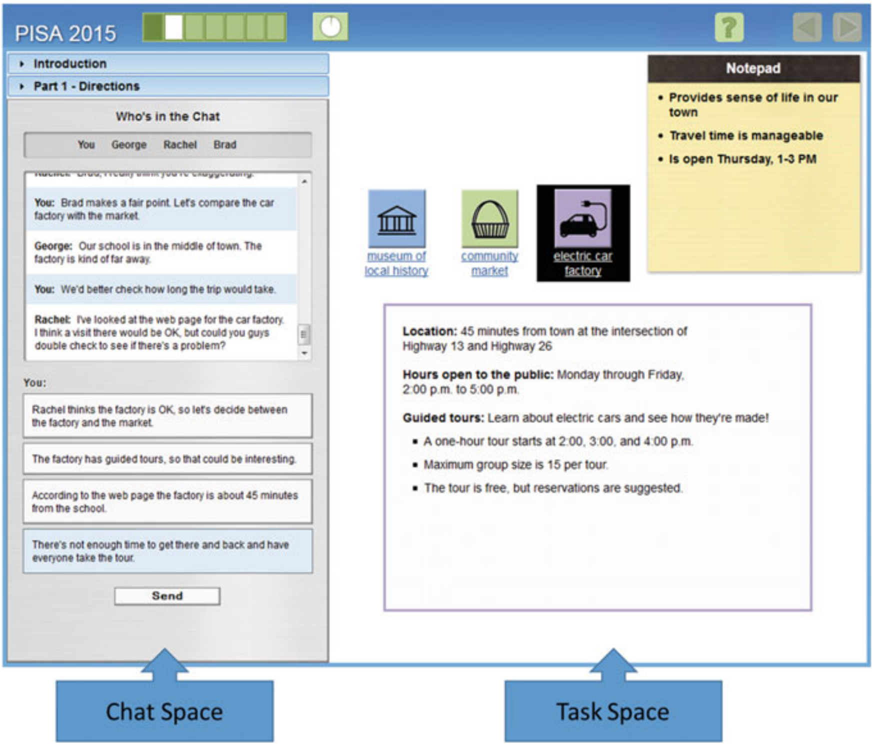
The CPS units included chat-based tasks where students interacted with one or more agents/simulated team members to solve a presented problem. Students were presented with a set of chat options and were asked to select the most appropriate statement in the "chat space" on the left side of the screen. Once selected, the choice displayed in the chat history area, and additional responses from one or more agents followed. Students could scroll through the history to review the chat as needed. Responses from agents were based on student selections. As a result, there could be multiple paths through each unit. To ensure that any incorrect or not-optimal selections would not penalize students as they progressed through the task, each unit was designed with convergence and rescue points (see more details in the subsequent section) to bring them back on task.

In addition to the chat interactions, the CPS units included a task area on the right side of the screen where students could take actions, view notes recorded by agents, or keep track of progress through the task. In the sample screen from part 1 of The Visit (Fig. 7.2), the "task space" included clickable links to three websites containing information needed to solve the problem assigned to the team as well as a notepad where teammates recorded key information.

7.2.3 Conversational Agent

In each CPS unit, the student worked with one or more group members to solve a problem, with the group members/computer agents providing input much as fellow students would do. The conversational agents responded to students' textual inputs and actions when the student moved through different stages of the problem. In each stage, communications or actions that could be performed by either the agent or the student were predefined, which resulted in the ability of objectively scoring all responses.

The computer dynamically monitored the state of the problem through the task completion process. Within each state, the students needed to carry on a conversation with the agent group members by making choices from a group of communication sets. Different students' responses may lead to different conversation paths or cause different actions from the agent as far as variations in the simulation or conversations. For instance, a conversational agent could add or reduce a task according to the student's choice, or respond to the student's request by providing an extra piece of information. Meanwhile, actions performed by the student during



Rachel: Brad, I really think you're exaggerating.

You: Brad makes a fair point. Let's compare the car factory with the market.

George: Our school is in the middle of town. The factory is kind of far away.

You: We'd better check how long the trip would take.

Rachel: I've looked at the web page for the car factory. I think a visit there would be OK, but could you guys double check to see if there's a problem?

You:

Rachel thinks the factory is OK, so let's decide between the factory and the market.

The factory has guided tours, so that could be interesting.

According to the web page the factory is about 45 minutes from the school.

There's not enough time to get there and back and have everyone take the tour.

Notepad

- Provides sense of life in our town
- Travel time is manageable
- Is open Thursday, 1-3 PM

Location: 45 minutes from town at the intersection of Highway 13 and Highway 26

Hours open to the public: Monday through Friday, 2:00 p.m. to 5:00 p.m.

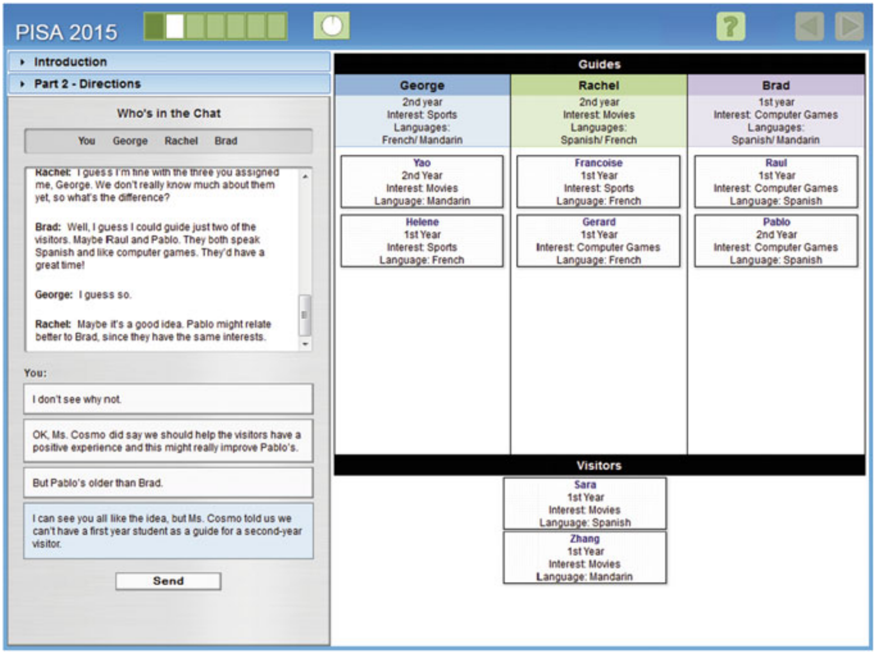
Guided tours: Learn about electric cars and see how they're made!

- A one-hour tour starts at 2:00, 3:00, and 4:00 p.m.
- Maximum group size is 15 per tour.
- The tour is free, but reservations are suggested.

Fig. 7.2 A sample screen of chat and task spaces in a released CPS item (The Visit) in PISA (OECD, 2015a)

the process of problem solving, such as moving an object and placing a time slot into a proposed schedule, were also monitored by the computer. The purpose of such monitoring was to track students' progress in task solving as well as record student actions related to the current stage of the problem (OECD, 2013).

Conversational agents can be utilized in various ways in a computer-based assessment, from simple chat interfaces to complex negotiations with multiple team members. In PISA 2015, the assessment of students' CPS skills was designed to take place in diverse environments, which allowed students to "work" with different agents and groups in order to cover the range of aspects defined in the CPS constructs. For example, in the released unit (The Visit), the student was required to supervise the work of agents where there is an asymmetry of roles, serving as a measurement of CPS skill D3 (monitoring, providing feedback, and adapting the team organization and roles). When an agent went off task a bit ("Who cares? All of these choices are boring. Let's take our visitors someplace they'll actually enjoy"), the credited response ("Brad, you're right that we want them to enjoy themselves, but we should discuss Ms. Cosmo's options first") acknowledged Brad's statement



Rachel: I guess I'm fine with the three you assigned me, George. We don't really know much about them yet, so what's the difference?

Brad: Well, I guess I could guide just two of the visitors. Maybe Raul and Pablo. They both speak Spanish and like computer games. They'd have a great time!

George: I guess so.

Rachel: Maybe it's a good idea. Pablo might relate better to Brad, since they have the same interests.

I don't see why not.

OK, Ms. Cosmo did say we should help the visitors have a positive experience and this might really improve Pablo's.

But Pablo's older than Brad.

I can see you all like the idea, but Ms. Cosmo told us we can't have a first year student as a guide for a second-year visitor.

Guides

George	Rachel	Brad
2nd year Interest: Sports Languages: French/ Mandarin	2nd year Interest: Movies Languages: Spanish/ French	1st year Interest: Computer Games Languages: Spanish/ Mandarin
Yao 2nd Year Interest: Movies Language: Mandarin	Francoise 1st Year Interest: Sports Language: French	Raul 1st Year Interest: Computer Games Language: Spanish
Helene 1st Year Interest: Sports Language: French	Gerard 1st Year Interest: Computer Games Language: French	Pablo 2nd Year Interest: Computer Games Language: Spanish

Visitors

Sara
1st Year
Interest: Movies
Language: Spanish

Zhang
1st Year
Interest: Movies
Language: Mandarin

Fig. 7.3 A sample screen of conversational agents and dynamic changes in task space in a released CPS item (The Visit) in PISA (OECD, 2015a)

while reminding him about the team's task, providing feedback to keep the discussion focused.

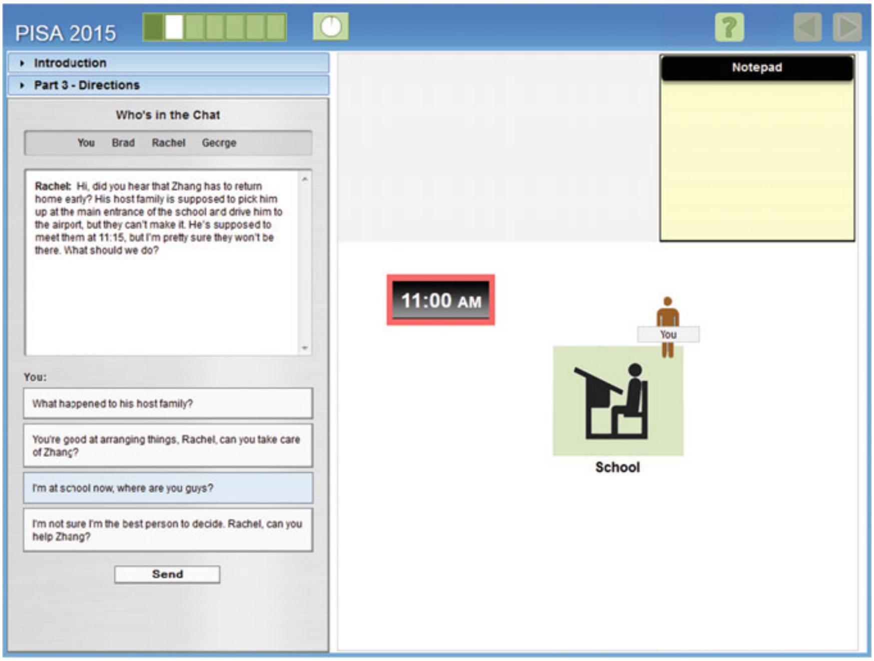
Other tasks involved disagreements between agents and the student regardless of whether the agent was collaboratively oriented (e.g., initiated ideas, supported and praised other team members) or not (e.g., interrupted, commented negatively about work of others). For instance, in the sample CPS unit, the two agents agreed about their tasks but had not met the teacher's requirement that a guide must be of equal or higher class rank than the visitors assigned to them. The student needed to remind the team to meet this requirement in order to gain credit (see Fig. 7.3).

7.2.4 *Convergence and Rescue Structures*

Convergence and rescue were the two main design concepts used in the PISA 2015 CPS units. Convergence was generally used to guarantee that different paths arrived at an identical point. That is, regardless of what choices the student made, the path led to the same convergence point. Each path to the convergence point had to provide the student with the same information and bring him or her to the same stage of the problem. The paths where students made noncredit or suboptimal choices generally had to incorporate rescue, the provision of required information through the agents or some other mechanism.

Many of the scenarios featured simple convergence and rescue structures. Units structured in this way began by presenting the student with a choice in the chat interface or an opportunity to perform a needed action in the task space. If the student failed to make the choice or take the action that advanced the team's progress toward solving the problem, an agent would do so in the next move. Then the student would be presented with another opportunity to display CPS behavior. Because students progressed through each scenario in a linear fashion, they could not go back or change their earlier responses. However, scrolling back and reading the chat record were allowed.

There were also some scenarios designed with more complex structures in which a student might have the possibility of going through two or three choice points before coming back to the convergence point. For example, at the beginning of part 3 in The Visit, the student and the agents needed to help one of the foreign students get to the airport (see Fig. 7.4). The full credited response was the third choice ("I'm at school, where are you guys?"), which told the team his or her location and led directly to the convergence point (i.e., exhibiting CPS skill B3—describing roles and team organization). But students who chose the alternative paths still arrived at the convergence point, although it took longer. For instance, if the student



Rachel: Hi, did you hear that Zhang has to return home early? His host family is supposed to pick him up at the main entrance of the school and drive him to the airport, but they can't make it. He's supposed to meet them at 11:15, but I'm pretty sure they won't be there. What should we do?

What happened to his host family?

You're good at arranging things, Rachel, can you take care of Zhang?

I'm at school now, where are you guys?

I'm not sure I'm the best person to decide. Rachel, can you help Zhang?

Fig. 7.4 A sample screen of convergence and rescue designs in a released CPS item (The Visit) in PISA (OECD, 2015a)

selected the first option (“What happened to his host family?”), Rachel rescued by saying she didn’t know what happened to his host family and asking the student if he or she were at school; this gave the student a second chance to choose the response providing his or her location. If the student selected the second option (“You’re good at arranging things, Rachel, can you take care of Zheng?”) or the fourth option (“I’m not sure I’m the best person to decide. Rachel, can you help Zheng?”), Rachel rescued by saying she is at home, with the student then being given the opportunity to respond by asking where everyone else is. The process data in the log file indicate that students were unlikely to notice these convergence and rescue structures. The structure design apparently made little impact on students’ test-taking behavior as they progressed through the scenario.

7.2.5 Measurement of CPS Skills

CPS is a conjoint process that combines problem solving and collaboration into one assessment domain. Through this complex process, students need to figure out the problem and find solutions as they interact with others, regulating social processes and exchanging information. How to make consistent, accurate, and reliable measurement in CPS across individuals and populations is a compelling question. This is a complex challenge when the collaborative interactions are set to occur in realistic environments (OECD, 2013).

Using computer-based agents provided the possibility to assess collaborative skills in an operationally feasible way to ensure standardized and comparable observations in large-scale assessments (OECD, 2013). This approach provided a high degree of control on the collaboration with conversational agents and standardization required for measurement, as well as flexibility for students to choose the optimal paths through the collaboration. Furthermore, it facilitated the PISA test administration by placing students in a variety of collaborative situations and allowing measurement within the time constraints.

The analyses for PISA CPS were conducted on the student level so that the design reflects measuring individual competencies rather than the overall performance of the process in which the teammates engage. Although the PISA 2015 CPS assessment was not designed to measure individuals’ cognitive problem-solving skills specifically, it featured a level of measurement of the individual problem-solving skills expressed through collaboration (OECD, 2015a). A complex data set was generated during the process of solving a problem in a collaborative situation in CBA, which included actions performed by the team members, communication among the group members, and products made by the individual and the group. Each item in the CPS domain can be associated with a level of proficiency for each CPS dimension. Because the focus is on the individual student as a representative of his or her country or subpopulation, measurement is on the output of the student. The output from the rest of the group provides contextual information about the state of the problem-solving process (OECD, 2013).

7.3 CPS Data Analysis in Field Trial PISA 2015

7.3.1 Item Response Theory (IRT)-Based Data Analysis

As described above, the PISA 2015 CPS units were based on simulated conversations with one or more computer-based agents that were designed to provide a virtual collaborative problem-solving situation. Test takers had to choose an optimal sentence from a multiple-choice list to go through the conversation with agents, or choose one or more actions programmed in the unit. Because of the similar item structures in other domains in PISA 2015, the data collected in the CPS units were evaluated by IRT models (Lord, 1980; Rasch, 1960) to establish reliable, valid, and comparable scales. The CPS scale in the main survey consists of six units, which in turn comprises multiple items within each unit that can be used for the IRT scaling. It was found that data from two units had dependencies in the responses due to different paths that could be taken by students through the simulated chat. Therefore, the CPS chat items that showed this kind of dependency were combined into composite items by summing the responses for the different paths that respondents could take. With this approach it was determined that each path-based response string could be scored to provide valid data and introduced into the IRT analysis. The composite items were used to generate polytomous items for the purpose of reducing issues with local dependencies.

All missing responses in the CPS domain were scored as not administered, because the administration of this domain required a response from each student at each stage, that is, students had to make a sequence of choices and could not skip forward. Those not-observed responses in CPS items were actually a result of students taking different paths while working on an item, meaning in the multiple-path situation, only one path could be taken, while other paths had to be missed. Therefore, not-observed responses do not reflect students' CPS skills and need to be treated as not administered. For the initial IRT analyses summarized here, the sample in the field trial was divided by country and language of administration, resulting in 55 country/language groups for CPS.[3]

For the new scales in the CPS assessment, a multigroup Rasch/partial credit model (PCM) and a multigroup two-parameter logistic model/generalized partial credit model (2PL/GPCM) were chosen as the scaling models. Each response at each stage of the unit was scored as either being indicative of CPS skills or not based on the scoring guide provided by the developers of the assessment units. A concurrent calibration was used to evaluate whether CPS items were functioning comparably across country and language groups or whether there were item-by-country or -language interactions. Item parameters for CPS items that were provided for the countries and used to identify items for the main study are based on

[3] The sample size for each country/language group was required to be 1950 students.

the 2PL/GPCM due to the improved model-data fit over the PCM model (see Table 7.1) and because more information (with regard to slope parameters) about each single item is provided. These item parameters were also used for generating a proxy score (expected a priori, or EAP) standardized within countries that is available in the data delivery to countries.

In order to examine the appropriateness of the IRT models, the item parameters for CPS items across countries, languages, and item fit statistics were calculated. For overall model fit, both Akaike information criteria (AIC; Akaike, 1974) and Bayesian information criteria (BIC; Schwarz, 1978) are provided. The item fit statistics used are the mean deviation (MD) and the root mean square deviation (RMSD). Both measures quantify the magnitude and direction of deviations in the observed data from the estimated item characteristic curve (ICC) for each single item. While the MD is most sensitive to the deviations of observed item difficulty parameters from the estimated ICC, the RMSD is sensitive to the deviations of both the observed item difficulty parameters and item discrimination parameters. In contrast to other measures that provide confidence intervals for the evaluation of model data fit, the MD and RMSD indices are not affected by sample-size issues that tend to result in many significant deviations observed in large samples when using Rasch-based infit and outfit measures. Moreover, MD and RMSD are available for a range of IRT models, while infit and outfit are typically only provided for the Rasch model.

The item fit of the CPS items was evaluated with regard to the concurrent calibration. The percentage of RMSD and MD was considered to be deviant using a rather strict criterion of RMSD > 0.20, and MD > 0.20 and < −0.20. It was found that item deviations for CPS items were generally small, only 0.9% and 0.5% of CPS items beyond the criterion of RMSD and MD, respectively. The deviation frequencies were not found to be substantially higher for any one particular country or language group. The results illustrate that the items show a good fit when using the same item parameters across different countries and languages. Moreover, the scale shows sufficient IRT-based (marginal) reliabilities (Sireci, Thissen, & Wainer, 1991; Wainer, Bradlow, & Wang, 2007) with 0.88 for CPS.

The specific structure of the CPS units and response types, as well as the results of the IRT analysis of the CPS using unidimensional models, prompted the need to conduct additional analyses. However, the unidimensional IRT models used in the assessment showed appropriate fit in terms of MD and RMSD (see the overall model fit in Table 7.1). Therefore, we were able to generate a standardized proxy score that could be used for verification of data collected within countries. This proxy is the EAP estimate, standardized within country, based on the unidimensional model with only international parameters. Initial item analyses with this CPS-based proxy as dependent variable were shared with countries. This score will be suitable for initial explorations of the associations of background variables with a quantity that reflects the common variance of collaborative skills assessed with the set of CPS items.

Table 7.1 Comparison of Rasch Model/PCM and 2PL Model/GPCM for CPS Items (OECD, 2015b)

	Likelihood	A-penalty	AIC	B-penalty	BIC
Rasch/PCM	−985,478	686	1,971,641	3877	1,974,832
2PL/GPCM	−971,209	994	1,943,411	5618	1,948,035

Note: *PCM* partial credit model; *2PL/GPCM* two-parameter logistic/partial credit model; *AIC* Akaike information criteria; *BIC* Bayesian information criteria

7.3.2 Correlations Between CPS Clusters

In the PISA 2015 field trial, CPS units were formed into four clusters for test administration. The correlation coefficients between clusters were generally reasonable, with a range from 0.76 to 0.81. The structure of the CPS units was such that there were a relatively large number of observables within a unit, while the number of units was small. The contextual coherence of the chat selections following a common theme within a unit may lead to the conjecture that what is measured is more the understanding of what a particular topic requires and may therefore be very specific to each unit.

7.4 Discussion

Collaboration is becoming increasingly important in the modern world as humans become more connected around the globe. The skill to efficiently solve a problem together with others is of special importance across educational settings and in the workforce. Compared with the problem-solving domain in PISA 2012 where the problem solving was defined as “individuals working alone on resolving problematic situations where a method of solution is not immediately obvious” (OECD, 2010, p. 12), the CPS domain in PISA 2015 broadens participation in problem solving from individuals to a group that is expected to join efforts and work together. Collaboration has distinct advantages over individual problem solving from at least two aspects: first, it allows for an effective division of labor by incorporating information from multiple sources of knowledge, perspectives, and experiences; second, it enhances creativity and quality of solutions stimulated by ideas of other group members (OECD, 2013). This chapter draws on measures of the CPS domain in PISA 2015 to address the development and implications of CPS items, challenges, and solutions related to item design, as well as computational models for CPS data analysis in large-scale assessments. Measuring CPS skills in PISA 2015 embraces both challenge and opportunity. On the one hand, it is a challenge compared to measuring individual skills alone; however, on the other hand, it makes observable the cognitive processes in which team members engage.

Regarding the importance of CPS measurement in large-scale assessments, some future work merits discussion. From the aspect of research methodologies, the analysis of CPS will continue in order to further explore how best to address and balance the between-unit versus within-unit variability. All items within a unit are most likely associated by the overarching topic of the simulated conversation more than the item responses given across units.

The use of aggregate scores will be further explored to allow for a definition of an overall level of adherence to the collaborative choices in the simulated conversations. These aggregate scores either could be provided by content experts who score and synthesize the major expected forms of behavior or empirically derived using latent class models that include order constraints, for example, the linear logistic latent class analysis (Formann, 1985, 1989, 1992).

A second approach is to come up with item attributes in order to analyze the units with respect only to items that belong to an attribute type. One possible distinction could be dividing items into ones that are predominantly communicative in nature and ones that are based more on the respondents' actions without a direct relation to a chat or communication with the simulated agent. It can be questioned whether, in a virtual environment, these actions are indeed considered by the respondent differently and would hence have potential to address different aspects of CPS. Whether this is indeed the case could be analyzed with additional multidimensional models that split items according to the "action" versus "chat response" attributes.

The CPS framework with computer agents was compatible with the capabilities of the PISA 2015 computer platform. The student could interact with the agents via a chat window, allowing the student to respond through communication menus. With respect to the student inputs, there were conventional interface components, such as mouse clicks, sliders for manipulating quantitative scales, drag and drop, cut and paste, and typed text input. Aside from communicating messages, the person could also perform actions on other interface components. For instance, additional data could be collected on whether students verified in the CPS environment whether actions been performed by an agent or whether they performed an action that the agent failed to perform. These actions are stored in a computer log file, which may provide additional information for tracking students' efforts in solving the CPS units.

Technical advances in computer-based learning systems have made greater efficiency possible by capturing more information about the problem-solving process. The availability of process log data sequences along with performance data has stimulated interest in education research and appears promising (e.g., Goldhammer et al., 2014; Graesser et al., 2004; Sonamthiang, Cercone, & Naruedomkul, 2007). For instance, He and von Davier (2016) drew on process data recorded in problem solving in technology-rich environments items in PIAAC to address how sequences of actions (n-grams) recorded in problem-solving items are related to task performance. Sukkarieh, von Davier, and Yamamoto (2012) used the longest common subsequence algorithms (LCS; e.g., Hirschberg, 1975, 1977) in a multilingual environment to compare sequences that test takers selected in a reading task against expert-generated ideal solutions. These methods are worth further exploration to investigate the associations between sequences of actions and CPS skills and to extract sequence patterns for different CPS proficiency levels.

In conclusion, PISA 2015 CPS competency is a conjoint dimension of collaboration skills, which serves as a leading strand, and problem-solving skills, which functions as an essential perspective. The effectiveness of CPS depends on the ability of group members to collaborate and prioritize the success of the group over that of the individual. At the same time, this ability is a trait in each of the individual members of the group (OECD, 2013).

This chapter looked at the CPS measures in PISA 2015, which was also the first trial of CPS units in a large-scale assessment. Besides giving a brief introduction to the development of CPS units in PISA, we used a sample unit to illustrate the structure of CPS items, challenges, and solutions related to item design, and the measurement of CPS skills in PISA. For future studies, we recommend using multivariate statistical analyses to address different aspects of CPS units and combining these analyses with process data from log files to track the process of students' learning and collaborative activities.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics, 118,* 1279–1333.

Cannon-Bowers, J. A., & Salas, E. (2001). Reflections on shared cognition. *Journal of Organizational Behavior, 22,* 195–202.

Cascio, W. F. (1995). Whither industrial and organizational psychology in a changing world of work? *American Psychologist, 50,* 928–939.

Dillenbourg, P. (Ed.) (1999). *Collaborative learning: Cognitive and computational approaches* (Advances in Learning and Instruction Series) New York, NY: Elsevier Science.

Dillenbourg, P., & Traum, D. (2006). Sharing solutions: Persistence and grounding in multi-modal collaborative problem solving. *Journal of the Learning Sciences, 15,* 121–151.

Fiore, S. M., & Schooler, J. W. (2004). Process mapping and shared cognition: Teamwork and the development of shared problem models. In E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 133–152). Washington, DC: American Psychological Association.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology, 38,* 87–111.

Formann, A. K. (1989). Constrained latent class models: Some further applications. *British Journal of Mathematical and Statistical Psychology, 42,* 37–54.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87,* 476–486.

Goldhammer, F., Naumann, J., Selter, A., Toth, K., Rolke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626.

Goos, M., Manning, A., & Salomons, A. (2009). Job polarization in Europe. *American Economic Review, 99,* 58–63.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers, 36,* 180–192.

Greiff, S., Wüstenberg, S., Csapo, B., Demetriou, A., Hautamäki, J., Graesser, A. C., et al. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review, 13,* 74–83.

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. Dordrecht, Netherlands: Springer.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749–776). Hershey, PA: Information Science Reference.

Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the ACM, 18,* 341–343.

Hirschberg, D. S. (1977). Algorithms for the longest common subsequence problem. *Journal of the ACM, 24*(4), 664–675.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, CA: Erlbaum.

Organisation for Economic Co-operation and Development. (2009). *PIAAC problem solving in technology-rich environments: A conceptual framework*. OECD Education Working Paper No. 36. Paris, France: Author.

Organisation for Economic Co-operation and Development (2010). *PISA 2012 Field Trial Problem Solving Framework.* Paris, France: Author. Accessed July 6, 2016 http://www.oecd.org/dataoecd/8/42/46962005.pdf

Organisation for Economic Co-operation and Development. (2013). *PISA 2015: Draft collaborative problem solving framework*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2015a). *PISA 2015 released field trial cognitive items*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2015b). *PISA 2015 field trial analysis report: Outcomes of the cognitive assessment (JT03371930).* Paris, France: Author.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Sonamthiang, S., Cercone, N., & Naruedomkul, K. (2007). Discovering hierarchical patterns of students' learning behavior in intelligent tutoring systems. In Institute of Electrical and Electronics Engineers (Ed.), *Proceedings of the 2007 IEEE International Conference on Granular Computing* (pp. 485–489). Los Alamitos, CA: IEEE Computer Society Press.

Sukkarieh, J. Z., von Davier, M., & Yamamoto, K. (2012). *From biology to education: Scoring and clustering multilingual text sequences and other sequential tasks* (Research Report No. RR-12-25). Princeton, NJ: Educational Testing Service.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving. More than reasoning? *Intelligence, 40,* 1–14.

Chapter 8
Assessing and Teaching 21st Century Skills: Collaborative Problem Solving as a Case Study

Patrick Griffin

Abstract This chapter describes the assessment of collaborative problem solving using human-to-human interaction. Tasks were designed to require partners to contribute resources or skills that they uniquely controlled. Issues were task design, data capture, item and data definition, calibration, and the link to teaching intervention. The interpretation of the student performance is mapped to a criterion-referenced interpretation framework, and reports are designed to assist teachers to intervene at a Vygotsky zone of proximal development in order to promote development of the student ability in collaborative problem solving. The data analytics demonstrate how the equivalent of test items are developed and issues such a local independence are discussed.

Keywords Collaborative problem solving · Human-to-human · Rasch modeling · Task design

8.1 Introduction

This chapter contributes an example of collaborative problem solving involving human-to-human interaction. As such, it complies with definitions offered by various writers and provides a model for developing collaborative problem-solving tasks. In doing so, many of the theoretical, technical, and practical issues are discussed and possibilities for future work on these areas are explored. The chapter explores how students interact in a problem-solving setting, how their actions and interactions are mapped, and how activity data in a digital environment are captured

An earlier version of this chapter was presented as a keynote lecture at the Institute of Curriculum & Instruction at East China Normal University, November 6–8, 2015, Shanghai, China.

P. Griffin (✉)
Melbourne Graduate School of Education, Parkville, Australia
e-mail: p.griffin@unimelb.edu.au

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_8

and used to identify patterns. Activity patterns are converted to a series of dichotomous and polytomous items for calibration and estimation of student ability. This also enables us to explore issues of dependence among partners and between items. While these issues are not resolved as yet, the chapter opens up a range of possible developments involving the measurement of groups as well as individuals in a collaborative setting. The implications for teaching are also explored, and the chapter provides reports likely to assist teachers in developing student ability in collaborative problem solving within a social constructivist model of learning.

8.2 Collaborative Problem Solving Measurement

Collaboration has become increasingly important in the 21st century. Bentley and Cazaly (2015) argued that collaboration is essential as it is increasingly sought after in education (as well as in other sectors) because it seems to offer key benefits such as efficient coordination of shared activities, authentic engagement, and relationships built through voluntary, reciprocal action, as well as flexible, differentiated support that matches teachers and learners with specific sources of support tailored to their specific needs and objectives (Dillenbourg & Traum, 2006; Fischer, Greif, & Funke, 2012; Kong, 2011; O'Neil, Chuang, & Chung, 2003; Organisation for Economic Co-operation and Development [OECD], 2013; Rummel & Spada, 2005).

Yet there is no consensus of definitions, and those definitions that are reported apparently depend upon the context in which the definition is offered. Nonetheless, some common language exists. The following definition has been compiled by putting together those characteristics that appear to be common across almost all definitions but add something that is missing from general definitions and discussions of collaboration. Collaboration is the sharing of effort, knowledge, and resources to pursue shared goals in ways that the collaborators cannot achieve alone, and there is a dependency among the collaborators who each must control and contribute unique resources in order to realize the shared goal.

The things that are missing reinforce the idea that collaboration is not based on a team of people, all of whom have the same skills, expertise, and resources. True collaboration brings together people who offer complementary skills, knowledge, materials, and other resources in order to understand and to build the joint understanding of the circumstances and realize a shared goal that they cannot achieve alone. First, there must be a shared and common goal; one must be able to analyze the situation, have a shared view of each person's unique role and contribution, and be willing to accept that no one person can solve the issue alone and that each partner depends upon another partner in order to proceed. So the missing characteristics are an inability to realize the goal alone, unique control of resources, and a dependency among participants.

8.2.1 What Is Driving Change?

Pressures for change and industry thirst for new ways of working, thinking, using tools, and creating lifestyle were identified by the Economist Magazine Intelligence Research Unit. In a study that included 26 countries and 19 different business sectors. Kenworthy and Kielstra (2015) identified four major issues that were putting pressure on education.

- Problem solving, team working, and communication are the skills that are currently most in demand in the workplace.
- Education systems are not providing enough of the skills that students and the workplace need.
- Some students are taking it into their own hands to make up for deficiencies within the education system.
- Technology has been changing teaching practice and resource use, but education systems are keeping up with the transformation rather than leading it.

This was consistent with the views of three major corporations which had, at the end of the 20th century, become concerned that education was not keeping pace with the changes of work and society. They argued that the knowledge, skills, attitudes, values, and ethics of the 21st century were undergoing fundamental changes compared to those of the 20th century. In a workshop that included a team of 250 experts, they identified four broad areas of skills needed in the workplace. They premised their discussions on the assertion that digital technology has changed the way we think, the way we work, the tools we use, and even the way we live and interact with others. The workshop participants wrote about new ways of thinking that included creativity and innovation, critical thinking, problem solving, decision making, learning to learn, and metacognition (Binkley et al., 2012).

The workshop participants examined the skills needed for new ways of working through communication and collaboration. They discussed the tools for working in the 21st century that required specific skills as well as information literacy and Information and Communication Technology. literacy. They raised our awareness that living in the world in the 21st century requires local and global skills of citizenship, flexible life and career skills, and an acceptance of personal and social responsibilities. This large group concluded that education needed to change quickly and fundamentally in order to cope with the pressures that digital technology was placing on working, living, employment, and even the way we think, because the control of information creation and distribution also influences what we think and what we know.

8.3 Collaboration

When we deal with collaboration and examine the way in which people work together, contributing their complementary skills, knowledge, resources, and experience in order to realize a shared goal of the group, the crossover to problem solving (Polya, 1973; Zoanetti & Griffin, 2014) becomes reasonably clear. A combination of collaboration, critical thinking, communication, and problem solving can be thought of as *collaborative problem solving* (CPS). The Assessment and Teaching of 21st Century Skills (ATC21S) project (Griffin, Care, & McGaw, 2012) set about defining ways of measuring individual person skills in collaborative problem solving.

From 2009 to 2012, the corporations Cisco, Microsoft, and Intel supported the development of a series of assessment tasks and teaching strategies that would enable schools to think about, entertain, and perhaps even implement 21st century skills assessment into their curricula. At the same time, the OECD, through its worldwide Programme for International Student Assessment (PISA) project, agreed to assess students in collaborative problem solving in the 2015 round of assessment. This meant that, potentially, CPS could be assessed in up to 65 countries, as a voluntary experimental measure in the 2015 PISA survey. During the period in which the assessment tasks were developed, it became clear that the programming language for this kind of work was shifting from Flash to HyperText Markup Language (HTML) 5. The University of Melbourne team had programmed everything in Flash, and the shift in technology meant everything had to be reprogrammed in HTML 5 but this gave the university an opportunity to improve, edit, and modify the tasks to make them more efficient. It also led to some fundamental breakthroughs in the way in which data were collected, coded, scored, and interpreted.

The Assessment Research Centre (Melbourne Graduate School of Education) explored new ways of interpreting a person's collaborative problem-solving skills. It became obvious very quickly that a classroom teacher could set up collaborative problem-solving tasks but would find assessing this work very difficult in a class full of students who were able to discuss, experiment, and communicate with one another while solving problems. It would be chaotic and impossible for a teacher to monitor and evaluate individual students in such a setting. The solution was to develop collaborative assessment tasks in a digital environment, such that monitoring and interpreting the students' work could be done electronically. This did not affect teaching the skills, because the teacher would still have classroom activities that enable collaboration when the time came to assess the students. The use of technology solved a very difficult classroom management problem for the teachers (Woods, Mountain, & Griffin, 2014).

8.4 Collaborative Problem Solving

Edwardo Salas (this volume) defined CPS as the situation where two or more individuals must interact and adapt to achieve specified shared and valued objectives. The ATC21S definition was more complex but consistent with the Salas definition. ATC21s combined critical thinking, problem solving, decision making, communication, and collaboration as CPS. Hesse, Care, Buder, Sassenberg, and Griffin (2014) argued that it consisted of a combination of social and cognitive skills. The social skills consisted of participation, perspective taking, and social regulation, and the cognitive skills consisted of task regulation and knowledge building. Participation skills involved action, interaction, and task completion or perseverance. Perspective-taking skills included elements of responsiveness to partners and audience awareness. Social regulation consisted of metamemory, transactive memory, negotiation skills, and responsibility initiative. The cognitive skills consisted of problem analysis, goal setting, resource management, and dealing with ambiguity. Many of these skills have been discussed by Salas, Von Davier, Graesser and others in this volume. The learning and knowledge-building skills were described as including data collection, systematicity, identifying relationships and patterns, explaining contingency or formulating rules, generalizing, and formulating hypotheses. Each of these were mapped through data analytics using an activity log file.

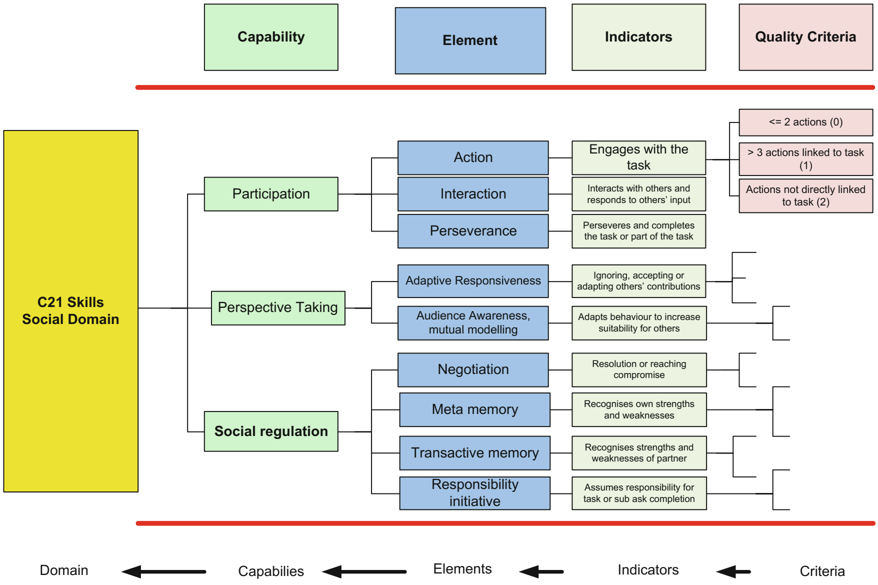
Art Graesser described the approach for PISA in 2015 using human-to-agent interaction. The PISA approach has a history of linking the work back to Polya's (1973) problem-solving framework, which was explored in PISA from 2003 to 2012 for problem-solving measurement, with the need to link collaborative problem solving in PISA to that history. ATC21S didn't have that constraint.

To illustrate what is meant by collaborative problem-solving tasks in ATC21S, we present a simple example. Suppose we have a jigsaw puzzle with 100 pieces. We randomly allocate 50 pieces to each of two students. The instruction to the students is, "Use these pieces of jigsaw puzzle to put the jigsaw together." Clearly, neither student could do this alone, because each one has only half the pieces. They must also understand how their actions and those of their partner can help to solve the overall puzzle. They also need to share the idea that the puzzle is solvable and that a strategy can be found in which they can each use their separate pieces of jigsaw to put the whole puzzle together. A critical moment occurs when one or both realize what is depicted in the puzzle, and then they are able to systematically test a range of ways to collaborate in solving the puzzle. There is dependency between the partners and an understanding that, by jointly working through the problem, it can be solved. Students attempting to solve CPS tasks typically carry out a preliminary start process of selecting and agreeing on roles with their partner. They explore and analyze the problem space, cooperate to identify resources each manages, and cooperate with partners in exploring and testing procedures leading to hypothesis formation and strategies to solve the problem. Some problem solvers then check whether any other solutions or strategies are possible.

The jigsaw puzzle example shows that each student controls specific resources (i.e., the jigsaw pieces) and they are able to contribute those resources uniquely to solve the problem. While leadership is important (the capacity to take responsibility), it should emerge as part of the collaborative problem-solving task requirements.

In developing a conceptual framework and hypothesized construct for CPS, a team of specialists in computer-supported collaborative learning (CSCL) and problem solving was assembled. The team was led by Hesse et al. (2014), and his team partitioned CPS into two major components—the social and cognitive components. Figure 8.1 through Sect. 8.4 illustrate how these components were defined using the format of rubrics illustrated by Griffin and Robertson (2014).

The social component consisted of three broad capabilities. These were the ability to participate, the ability to take the perspective of another person, and the ability to be able to understand the way in which the collaborative group members interacted and worked with each other. Participation could be further broken down into the actions taken by the student, interactions between the student and the partner, and the extent to which an individual would persevere and participate in the realization of the shared goal. Perspective taking could be seen as the extent to which the person was responsive to the actions and interactions of his or her partner and was able to adapt his or her own behavior in relationship to this. It was also argued by Hesse and his team that perspective taking involved the extent to which a person's actions and contributions were a result of their awareness of their partner's process. The social regulation or social organization of the group also led to the idea



Capability
Element
Indicators
Quality Criteria
C21 Skills Social Domain
Participation
Perspective Taking
Social regulation
Action
Interaction
Perseverance
Adaptive Responsiveness
Audience Awareness, mutual modelling
Negotiation
Meta memory
Transactive memory
Responsibility initiative
Engages with the task
Interacts with others and responds to others' input
Perseveres and completes the task or part of the task
Ignoring, accepting or adapting others' contributions
Adapts behaviour to increase suitability for others
Resolution or reaching compromise
Recognises own strengths and weaknesses
Recognises strengths and weaknesses of partner
Assumes responsibility for task or sub ask completion
<= 2 actions (0)
> 3 actions linked to task (1)
Actions not directly linked to task (2)
Domain
Capabilies
Elements
Indicators
Criteria

Fig. 8.1 The social component of CPS

of students' ability in negotiation, in analyzing their own strengths and weaknesses and adapting their contribution accordingly, in recognizing the strengths and weaknesses of their partners and adapting their own behavior, and finally their ability to accept responsibility and initiative or to show leadership within the group. Each of the elements and indicative behaviors was proposed such that, if evidence of the nominated behavior were to be observed, it would be possible to build a case for the presence of this particular element or skill in the person's repertoire; the quality criteria would indicate how much is present. In order to keep diagrams simple, only one indicative behavior is listed for each element. Of course, there would be several indicative behaviors for each of the elements. Each indicative behavior was then reviewed to define levels of quality with which that behavior could be exhibited, and these were organized as ordered criteria within a scoring rubric (see Griffin & Robertson, 2014). In the jigsaw puzzle example, the cognitive activity may be engagement with the task. A student might, for example, simply pick up individual pieces and, working alone, try to locate the best place for each piece. Others might sort the jigsaw pieces they've been given according to color, shape, and pattern. Others might encourage their partner to do the same and together explore different ways of assembling the pieces of the puzzle. What we can see in this is a hierarchy of quality of performance. These hierarchies were labeled *quality criteria* (Griffin & Robertson, 2014) (Fig. 8.2).

In the cognitive domain, a series of broad general capabilities consisted of the capacity of the student to analyze the task and to build knowledge through problem-solving behaviors. The elements associated with task analysis or task

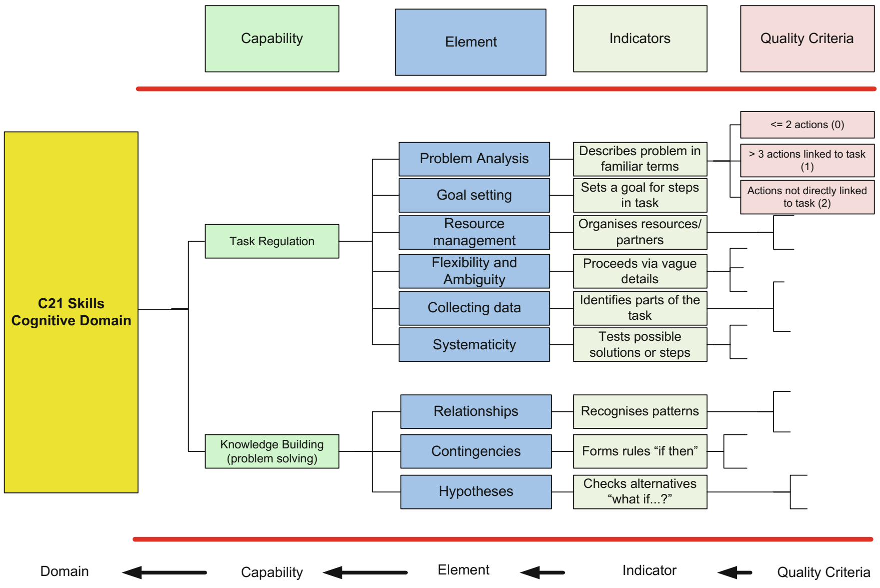

Capability
Element
Indicators
Quality Criteria
C21 Skills Cognitive Domain
Task Regulation
Knowledge Building (problem solving)
Problem Analysis
Goal setting
Resource management
Flexibility and Ambiguity
Collecting data
Systematicity
Relationships
Contingencies
Hypotheses
Describes problem in familiar terms
Sets a goal for steps in task
Organises resources/ partners
Proceeds via vague details
Identifies parts of the task
Tests possible solutions or steps
Recognises patterns
Forms rules "if then"
Checks alternatives "what if...?"
<= 2 actions (0)
> 3 actions linked to task (1)
Actions not directly linked to task (2)
Domain
Capability
Element
Indicator
Quality Criteria

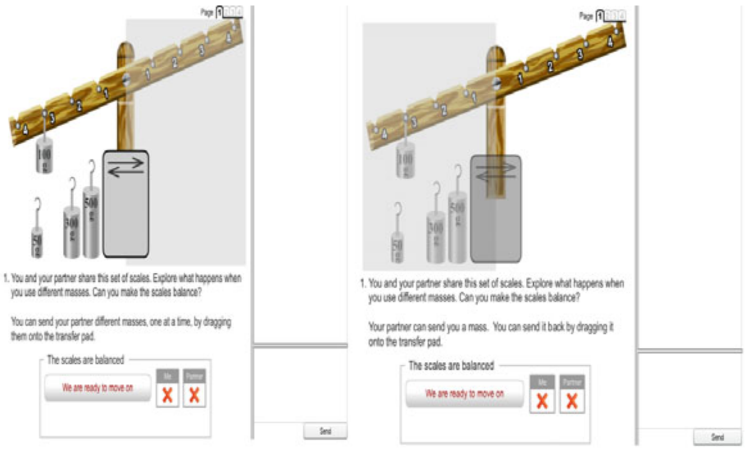
Fig. 8.2 The cognitive component of CPS

regulation were listed by Hesse et al. (2014) as the ability to analyze the problem, to set goals, to manage the resources each student can control, to be flexible in the way in which they go about this, and to be able to adjust and deal with confusion and ambiguity in the situation. Students might also collect data in order to make decisions systematically and to find a way to work through the task. The knowledge-building components of the cognitive domain involved the students' ability to identify patterns in their own and their partner's actions, especially the relationships between their own behaviors and those of their partner. They also need to examine patterns of activity between components of the task, their capacity to deal with contingency (that is, if Event A occurs, what will be the subsequent event?), and their capacity to formulate and test hypotheses. In constructing the assessments to measure these skills, each task solution contained a critical moment whereby one or more of the partners gained a sudden insight into the solution. Once that critical piece of information had been identified, it was possible for the partners to formulate a plan to proceed with solving the problem. This led to the idea of hypothesis formulation (Griffin, 2014) which could contain the phrases "What if…," "What about…," and "It seems to depend on…" in the discussion between students. These three basic expressions indicate that the student is entertaining possible alternatives regarding how to proceed. Moreover, these possible alternatives need to be tested, and patterns need to be identified within parts of the problem and hypotheses tested according to the queries made by the students. Let's examine two examples.

8.4.1 The Beam Balance

In this task, two students work together to place weights on the beam balance in order to bring the balance to equilibrium. Student A controls the weights and can transfer weights to student B. Student B makes the decisions about where to place the weights on the beam. Student B has four choices of location in which to hang each weight passed by Student A. The students can communicate using the chat box on the right of each screen. In this way, the communications are captured by the computer in a log file together with all other logged activity and chat data. The data analytics then provide an opportunity to explore cohesive patterns within logged chats and actions and, using a measurement model, to determine whether the patterns can be interpreted in terms of the relationship between behaviors in the cognitive and social conceptual frameworks (Fig. 8.3).

In a real life example, two children were playing on a swing in a playground. The swing is constructed as an equivalent of a beam balance with a seat for a child on each end of a bar. The children have to work out how to get onto the swing and balance it so they can both rotate and bounce up and down. The activity and participation of each child is clear. They both tried to solve the problem, initially on their own, by directing each other regarding what to do. Then one approached the other and suggested a common or shared approach. They realized that they must



1. You and your partner share this set of scales. Explore what happens when you use different masses. Can you make the scales balance?

You can send your partner different masses, one at a time, by dragging them onto the transfer pad.

1. You and your partner share this set of scales. Explore what happens when you use different masses. Can you make the scales balance?

Your partner can send you a mass. You can send it back by dragging it onto the transfer pad.

Student A view

Student B view

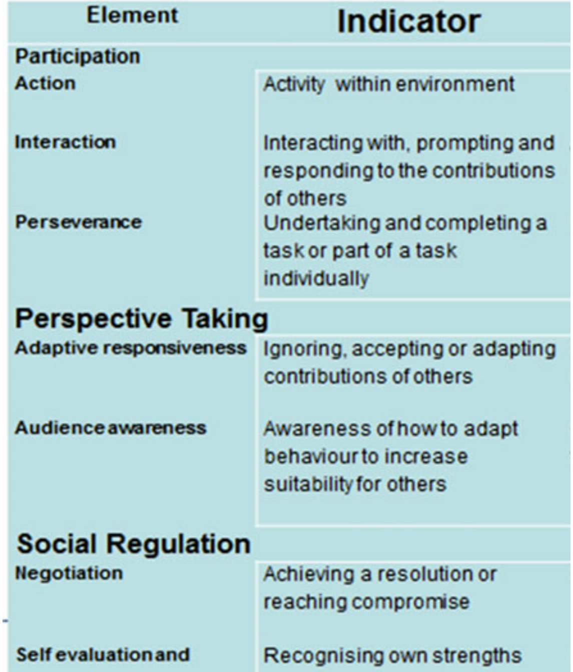
Fig. 8.3 Screen views for beam balance task

cooperate and make adjustments for the differences in their weights. They discussed at some length how they might go about this task. They persevered with a joint solution until they both were on the swing rotating and moving up and down. That is, they shared a common goal and developed a strategy for each person to implement in order to solve the problem and to realize their goal. They were also able to understand why the other person was having difficulty. This is evidence of their perspective-taking skills. They responded to and adapted their own behavior in order to accommodate the difficulties the other person was having. They were aware of the kinds of things that each was trying to achieve. They negotiated and explained to one another what difficulties they were having and what steps should be taken to overcome those difficulties. So the indicators literally came to life. In a video version of this task which runs for about two minutes, all of these indicators are evident. It is a simple task for the observer to recognize the indicators in a single episode, but in a class full of students all trying to work out the beam balance and discussing their motives and strategies, it would be very difficult for the teacher. The example can be found on Youtube at https://youtu.be/fwT7qI1ASfk. The elements and the behaviors elicited are described in Fig. 8.4.

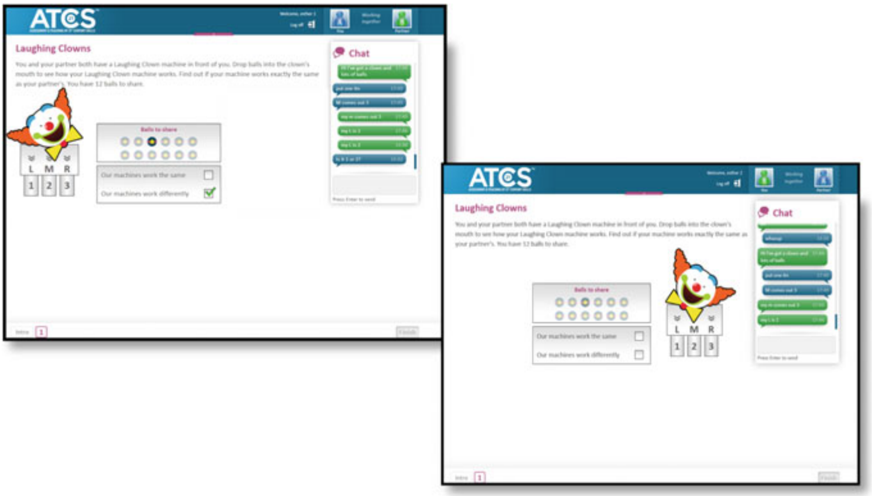
8.4.2 *The Laughing Clowns*

In ATC21S all tasks were administered to pairs of students, involving human-to-human interaction communicating via a chat box. Details of the design, implementation, and scoring of the tasks are provided by Griffin and Care (2014). In this collaborative task (Care, Griffin, Scoular, Awwal, & Zoanetti, 2014) two

Fig. 8.4 Elements and indicators of a CPS balance task in real life



Element	Indicator
Participation	
Action	Activity within environment
Interaction	Interacting with, prompting and responding to the contributions of others
Perseverance	Undertaking and completing a task or part of a task individually
Perspective Taking	
Adaptive responsiveness	Ignoring, accepting or adapting contributions of others
Audience awareness	Awareness of how to adapt behaviour to increase suitability for others
Social Regulation	
Negotiation	Achieving a resolution or reaching compromise
Self evaluation and	Recognising own strengths



Laughing Clowns

You and your partner both have a Laughing Clown machine in front of you. Drop balls into the clown's mouth to see how your Laughing Clown machine works. Find out if your machine works exactly the same as your partner's. You have 12 balls to share.

Balls to share

Our machines work the same ☐
Our machines work differently ☑

Laughing Clowns

You and your partner both have a Laughing Clown machine in front of you. Drop balls into the clown's mouth to see how your Laughing Clown machine works. Find out if your machine works exactly the same as your partner's. You have 12 balls to share.

Balls to share

Our machines work the same ☐
Our machines work differently ☐

Fig. 8.5 The fairground laughing clowns task

students see much the same thing, as shown in the example in Fig. 8.5. The screen that student B can see is a mirror image of the screen that Student A can see. The students are given 12 tokens (balls) to place in the clown's mouth to determine the relationship between the position of the chute when the ball is dropped into the clown's mouth and the position of the chute when the ball emerges. Their task is to see whether the input clown mouth leads to the same output location across the two clowns. Neither can see the other student's screen. The task can look simple but be difficult, because the students have to share the 12 tokens. Once a token is chosen and used by one student, it is not available to the other student. Consequently, skills of negotiation, communication, perspective taking, and participation have to be employed by each of the students. They need to adapt and monitor their own behavior and suggest modification to the behavior of their colleague so they can make adjustments of their own and the partner's understanding of the task. This leads again to the issue of local independence. As von Davier remarked (this volume) local independence means that a student's response to any item is independent of that student's response to any other item, given the underlying trait. Part of the problem of dependence related to CPS is understanding what is meant by an item.

8.5 Developing CPS Items

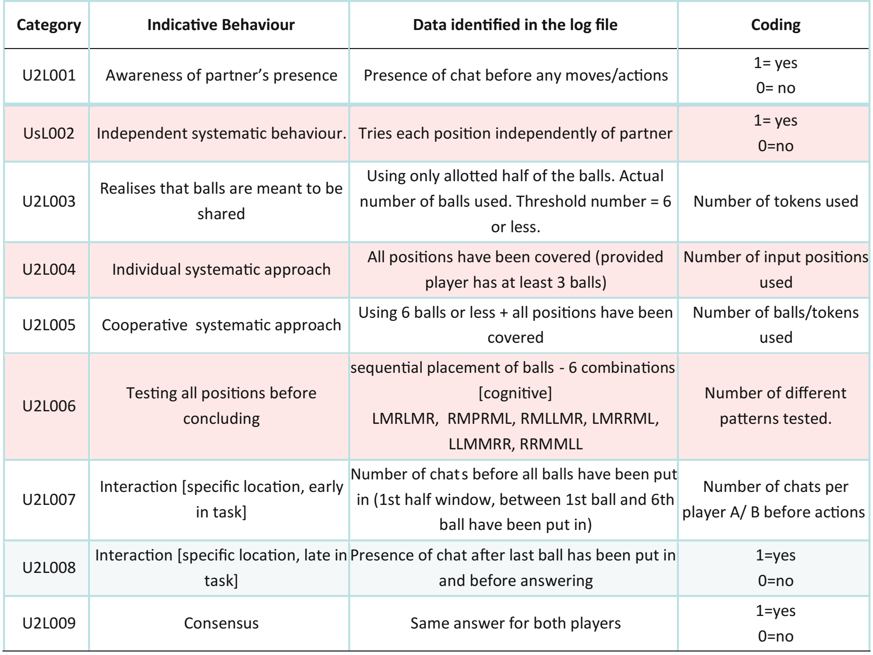
While the students are working through the assessment task, all of the actions, cursor movements, communications, and hesitations are monitored, logged, and timed and captured in a log stream data file. Log stream (sometimes called the click stream data) generated by a CPS platform includes digital traces of every action taken by every student in every part of the CPS task. CPS log stream contains a time-stamped record of each interaction of each student with each element of the platform. A record is generated when students log in, explore the problem space, communicate and interact with their partner, try different approaches and evaluate solutions. These are generally manifested by cursor movement, chat between partners, drop and drag, and so on. Every action is recorded and time stamped. These data provide considerable scope for investigation of patterns of problem solving-related collaborative behavior, and offer several advantages to such use. The coded data are digital traces of the interaction of the learner with the CPS platform. It is also possible to collect records or codes pertaining to each of the elements of the CPS as defined by Hesse et al. (2014).

Item response modeling and, in particular, the Rasch family of models offers one approach to interpretation of a student's CPS ability. These models estimate the probability of success by a person on a specific behavioral indicator or in terms of a test item, given the relative positions of the person and the item on an underlying variable or construct. When this is applied to CPS data, there has to be some cleaning of the data, classification and categorization of data points in the activity log file, and coding of behavioral data patterns such that each category of student CPS behavior in the log stream data becomes the CPS equivalent of a test item normally encountered in assessments of learning. Hence we can record for each CPS participant whether the category of data was present or absent (the equivalent of correct or incorrect). Each category that is recorded as present or absent for each student is named (given a code), and each coded category becomes the equivalent of a test item, but it consists of a complex set of behavioral file data. These will now be described as items but remembering that in this context, an item is a behavioral pattern arising from the interaction of the CPS student with the stimulus materials, partner, or context of the CPS platform. The presence or absence of these items (behavioral patterns) is then coded as *present* (1) or *absent* (0) using Adams, Vista, Awwal, Scoular, and Griffin's (2014) insight that frequency is a proxy measure of difficulty. In some cases, a polytomous record is used to indicate how much or how well an activity represents the element of the framework. The complexity of the data categories being renamed or coded as items makes the issue of local independence even more difficult to identify and manage.

The behaviors are recorded as shown in Table 8.1 and described in the right-hand column. The example shows how direct observation of manifest behavior can be used to infer the presence of the latent elemental skill. In analyzing these data, the frequencies of entries in the behavior categories were interpreted as relative difficulty estimates based on the assumption that the frequency with which a particular behavior occurred gave an estimate of the relative difficulty of that behavior (Adams et al., 2014). This was a reasonable assumption. If it were applied to the scoring and coding of multiple-choice test items, for example, a student's response would usually be recorded as 1 for the correct answer and 0 for the incorrect answer. Of course, the code 1 should be interpreted in terms of what student manifest behavior the record represents. If most students were able to select the correct answer, the frequency of the code 1 would be high. If most students were unable to find the correct answer, the frequency of the code 1 would be low.

Table 8.1 Examples of social elements and indicative behaviors for the laughing clowns task

Social element	Indicative social behavior	Example data used as evidence
Interaction	Interacts with partner	Presence of chat before allowing partner to make a move
Audience awareness	Adapts contributions to increase understanding for partner	Number of ball moves attempted before stopping and waiting for partner to move or respond
Responsibility initiative	Takes responsibility for progress for the group task	Number of times communicated with partner before the first half of the shared balls were used
Resource management	Manages resources	Realizes that balls are meant to be shared and uses only allotted half
Systematicity	Implements possible solutions to a problem	Uses the allotted half of the balls to cover the positions in a sequential order
Relationships	Identifies connections and patterns between elements of knowledge	Both students coming to an agreement on how their machine works
Solution	Arrives at correct answer	Selection of the correct option by A and B on how their machine works



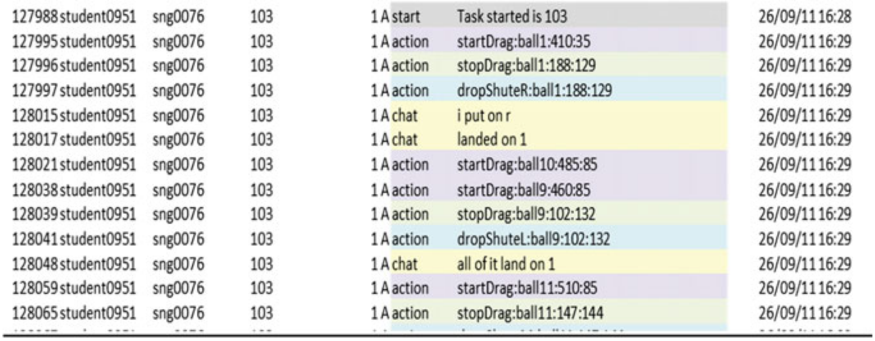
Category	Indicative Behaviour	Data identified in the log file	Coding
U2L001	Awareness of partner's presence	Presence of chat before any moves/actions	1= yes 0= no
UsL002	Independent systematic behaviour.	Tries each position independently of partner	1= yes 0=no
U2L003	Realises that balls are meant to be shared	Using only allotted half of the balls. Actual number of balls used. Threshold number = 6 or less.	Number of tokens used
U2L004	Individual systematic approach	All positions have been covered (provided player has at least 3 balls)	Number of input positions used
U2L005	Cooperative systematic approach	Using 6 balls or less + all positions have been covered	Number of balls/tokens used
U2L006	Testing all positions before concluding	sequential placement of balls - 6 combinations [cognitive] LMRLMR, RMPRML, RMLLMR, LMRRML, LLMMRR, RRMMLL	Number of different patterns tested.
U2L007	Interaction [specific location, early in task]	Number of chats before all balls have been put in (1st half window, between 1st ball and 6th ball have been put in)	Number of chats per player A/ B before actions
U2L008	Interaction [specific location, late in task]	Presence of chat after last ball has been put in and before answering	1=yes 0=no
U2L009	Consensus	Same answer for both players	1=yes 0=no

Fig. 8.6 Establishing codes and variable names based upon evidence in the log file

Items with large numbers of records coded 1 would be considered to be easy items, and items with low frequency of the code 1 would be considered difficult. This being the case, it enabled us to introduce a scoring process that could be analyzed according to measurement theory. In this case, we applied the Rasch model. Some partial credit data were also derived from the activity log files; these were mainly linked to time lapses and to repetitive actions (e.g., the number of chat exchanges). Details of the coding, scoring, and calibration are provided by Adams et al. (2014), and examples are provided in Fig. 8.6.

In terms of resource management, the student eventually has to realize that the tokens are meant to be shared and not used exclusively by them. Evidence of this might be that they used only half of the balls. So the number of balls used by an individual would be six or less. This would indicate that the student has realized that sharing and negotiation are mandatory. This realization or behavior category is given a name which becomes a way of recording each student's behavior. In this case, the name assigned to the behavior category is U2L003. Each part of the code has a meaning for later analysis, but details of that meaning are not necessary here. For each student, the number of tokens used is counted and recorded in this behavior category called U2L003.

A sample of the actual log file for the laughing clowns task undertaken by a Singaporean student is shown in Fig. 8.7. Given this method of recording it is a simple matter to count the number of tokens used by student 0951 from Singapore



```
127988 student0951   sng0076   103   1 A start    Task started is 103             26/09/11 16:28
127995 student0951   sng0076   103   1 A action   startDrag:ball1:410:35          26/09/11 16:29
127996 student0951   sng0076   103   1 A action   stopDrag:ball1:188:129          26/09/11 16:29
127997 student0951   sng0076   103   1 A action   dropShuteR:ball1:188:129        26/09/11 16:29
128015 student0951   sng0076   103   1 A chat     i put on r                      26/09/11 16:29
128017 student0951   sng0076   103   1 A chat     landed on 1                     26/09/11 16:29
128021 student0951   sng0076   103   1 A action   startDrag:ball10:485:85         26/09/11 16:29
128038 student0951   sng0076   103   1 A action   startDrag:ball9:460:85          26/09/11 16:29
128039 student0951   sng0076   103   1 A action   stopDrag:ball9:102:132          26/09/11 16:29
128041 student0951   sng0076   103   1 A action   dropShuteL:ball9:102:132        26/09/11 16:29
128048 student0951   sng0076   103   1 A chat     all of it land on 1             26/09/11 16:29
128059 student0951   sng0076   103   1 A action   startDrag:ball11:510:85         26/09/11 16:29
128065 student0951   sng0076   103   1 A action   stopDrag:ball11:147:144         26/09/11 16:29
```

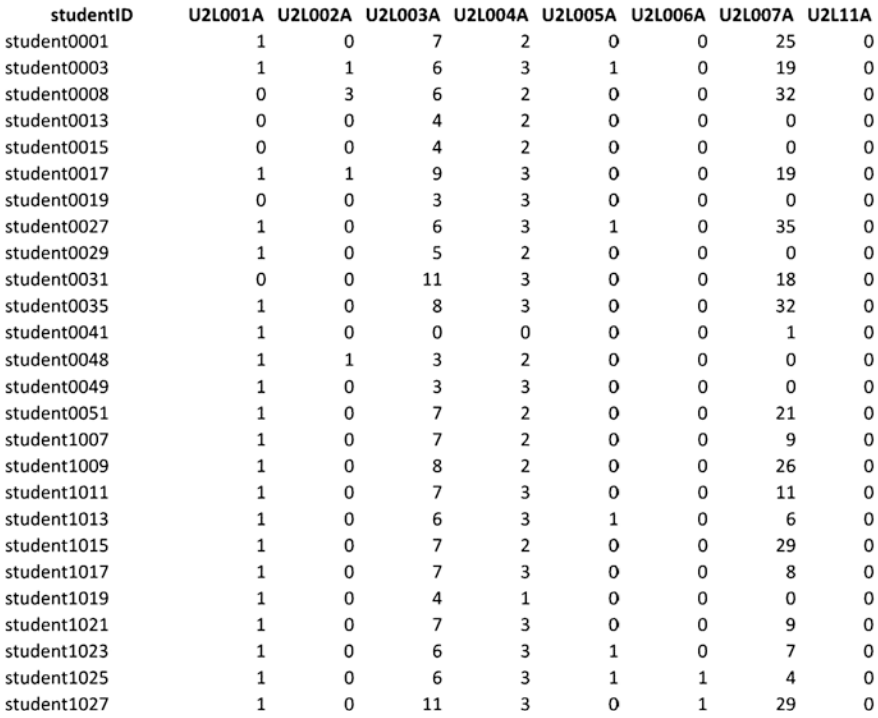
Fig. 8.7 Sample activity log file for the laughing clowns task

working on task 103 (laughing clowns) as Student A. The coding procedure also demonstrates the action of dragging a ball (token) from the location of the set of tokens (410:35) to a location whose coordinates (188:129) can be translated into clown's mouth location.

In Fig. 8.7, columns identify the student, the country, the task, whether this is Student A or B, the action class (communication or activity), the data, and the timing. It can be seen in the data that the student is dragging and dropping the tokens, and it is also clear regarding the number of tokens that are being dragged. This activity log file records every move, every communication, and every action that the student undertakes. It also separates the actions of Student A from those of Student B. Counts of the actions taken over all students in role of Student A or Student B enable frequencies of behaviors to be logged, and the relative difficulty can then be estimated using a measurement model to calibrate the task and to estimate the ability of this student (0951) acting in the role of Student A, independent of the estimation of Student B's ability based in the presence or absence of behaviors in each category. A separate estimation analysis of Student B's behavior can be undertaken. This suggests that methods proposed by Chow and others (this volume) in their dynamic exchange model might provide opportunities to progress, but the examples of mother-infant interaction represent total interpersonal dependence, and the ATC21S project went to some trouble to avoid this source of disturbance.

Examples of the coding and evidence data can be seen in Fig. 8.6. Each individual data piece is identified in the log file and recorded in a similar manner as that used to code and record the realization evidence illustrated above.

For each student, the presence or absence of each datum, or a count of relevant events found in the log file, is recorded for each of the categories (now called arrays), yielding a data file suitable for further analyses, and in particular the imposition of a measurement model searching for a coherent set of categories that could be interpreted in terms of the defined construct underpinning the behavior (ability) of the student.



studentID	U2L001A	U2L002A	U2L003A	U2L004A	U2L005A	U2L006A	U2L007A	U2L11A
student0001	1	0	7	2	0	0	25	0
student0003	1	1	6	3	1	0	19	0
student0008	0	3	6	2	0	0	32	0
student0013	0	0	4	2	0	0	0	0
student0015	0	0	4	2	0	0	0	0
student0017	1	1	9	3	0	0	19	0
student0019	0	0	3	3	0	0	0	0
student0027	1	0	6	3	1	0	35	0
student0029	1	0	5	2	0	0	0	0
student0031	0	0	11	3	0	0	18	0
student0035	1	0	8	3	0	0	32	0
student0041	1	0	0	0	0	0	1	0
student0048	1	1	3	2	0	0	0	0
student0049	1	0	3	3	0	0	0	0
student0051	1	0	7	2	0	0	21	0
student1007	1	0	7	2	0	0	9	0
student1009	1	0	8	2	0	0	26	0
student1011	1	0	7	3	0	0	11	0
student1013	1	0	6	3	1	0	6	0
student1015	1	0	7	2	0	0	29	0
student1017	1	0	7	3	0	0	8	0
student1019	1	0	4	1	0	0	0	0
student1021	1	0	7	3	0	0	9	0
student1023	1	0	6	3	1	0	7	0
student1025	1	0	6	3	1	1	4	0
student1027	1	0	11	3	0	1	29	0

Fig. 8.8 Sample data structure for students participating as student A

The structure of the data file is illustrated in Fig. 8.8. Each of the vertical columns represents a data array used to record evidence of the student behavior on each of the coded categories. It can be seen that for U2L001A the records are either 1 or 0 to indicate presence or absence for a student. A code of 1 is recorded if the student was aware of the partner's involvement and took action accordingly or 0 if he or she did not. For the variable U2L003A, the numbers recorded indicate the number of tokens that Student A has used. This enabled further recording using the threshold of 6 to indicate whether or not this student had realized that sharing was possible and even mandatory in order to realize the shared goal. All item data are reported in Fig. 8.8.

For the variable U2L006A, Student A's behavior is recorded to indicate whether all possible combinations of exit chute were tested. A code of 1 indicates that this was done. A code of 0 indicates that it was not done. Very few students tested all possible combinations, and this was interpreted as a very difficult behavior to exhibit. Awareness of the partner's presence and involvement was easy to exhibit, and most of the records represent that with a score of 1. Analyzing these data using an item response model (Rasch, 1960/1980) enabled the relative difficulties of behavior categories to be associated with the relative abilities of each of the students. High ability students are represented by the Xs at the top of the distribution in

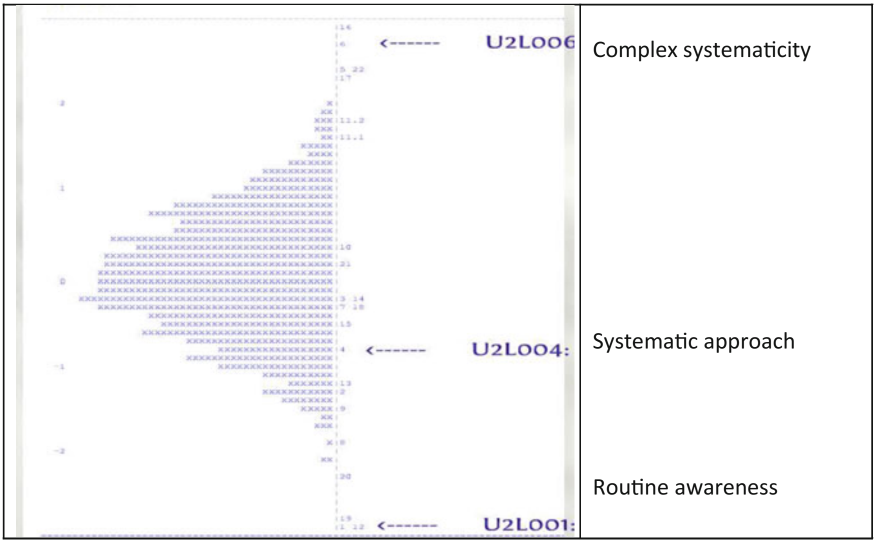


Fig. 8.9 IRT analysis of the indicative behaviors interpreting the construct

Fig. 8.9, which is called a *Wright map* (Wilson, 2009). Those students are likely to demonstrate the high-level systematicity for behavior category U2L006A and almost certainly exhibit the lower difficulty behavior indicated by their responsiveness to the presence of a partner in behavior category U2L001A. The distribution of student abilities indicates that the medium difficulty behavior, covering all of the positions, designated by variable U2L004A, was relatively easy and demonstrated by many students. It is of medium difficulty. Each of these data points in the activity log file became the equivalent of a dichotomous or polytomous test item to be coded, scored, and included in the calibration and estimation of student ability.

Concern about loss of local independence was considered. However, based on the work of Verhelst and Verstralen (1997), a test to overcome the loss of local independence would be to model the full set of items as one partial credit item. The dependency would then be taken into account. However, it would then not be possible to match the item parameters to individual items in the set. The loss of information pertinent to teaching of 21st century skills was deemed to be the greater evil. The decision was based in part on the lack of substantial evidence that the local independence of items was violated. The data were nonpredictable and not Guttmann-like (Baghaei, 2007), and given the conservative reporting of student ability (linked to teaching intervention), it was decided to proceed as if local independence was not violated. If the project aimed to estimate population parameters, this issue would warrant further examination.

By placing the descriptions of the behavior on the right-hand side of the chart as an interpretation of the behavior category, it was possible to get a sense of the relative development or growth characteristics of students distributed over the range of the construct. Students with greater amounts of the collaborative problem-solving skill would be at the top of the distribution, exhibiting very complex systematic behavior, and students with very little of the collaborative problem-solving ability would be at the bottom of the distribution, exhibiting routine awareness of the problem or their partner. Much more complex descriptions of the behavior would therefore be possible by interpreting each of the codes on the right-hand side of the item response variable map.

In order to make this meaningful in reports issued to both students and teachers, an interpretation of the construct is partitioned into levels of increasing proficiency or competence (Glaser, 1983). Brief descriptions of these levels are given to both the student and the teacher in a series of reports that indicate the level of the progression reached by the student. This level reached is set where the student can demonstrate approximately 50% of the behavior categories or where the student has approximately a 50% chance of being able to demonstrate the set of behaviors clustered at that level. This 50% chance enables us to link this to the learning theory of Vygotsky (1978) as the student's zone of proximal development. It demonstrates the level at which the student is most ready to learn with assistance.

In order to optimize the information provided to the teacher, collaborative problem solving tasks were 'bundled' to ensure that each bundle could provide adequate information regarding the student's development progress in collaborative problem solving.

Three reports were issued as illustrated in Fig. 8.10. The first was a Learning Readiness Report for each of the components in collaborative problem solving. This report provides the teacher with an estimate for each student of the point of intervention where learning is most likely to be promoted (the zone of proximal development). This approach to reporting was prompted by the formative assessment approach of the ATC21S project, where teachers were expected to be given advice which could be used to help students develop their skills in collaborative problem solving. A posttest result, when superimposed on the report, would indicate how much the student has progressed and the nature of that growth. The second report was a class profile report indicating where each student in the class had reached on the progression and the kinds of interventions that would promote

Fig. 8.10 Reporting to teachers and students

student growth. The chart illustrated the relative levels of student development within the class. The report is provided to the teacher for purposes of flexible classroom management and instructional grouping of students. Each class report can also accommodate a second or posttest report, which indicates to the teacher the amount of growth for each student, the nature of the growth, the rate of growth, as well as the rate of growth of each student in the class relative to other students, both from the same baseline (value add) and from various baseline measures of all students in the class. The third report is a student profile report which indicates the relative development or growth across each of the five strands of the dimensions of collaborative problem-solving (participation, perspective, group work, task analysis, and problem solving skills). The profile report also contains dates of each assessment and enables a second or posttest performance to be mapped. In so doing, the report indicates the amount and rate of growth. Furthermore, these reports are available to the teacher and the student within seconds after the students complete the assessments.

8.6 Differential Test Functioning

The project was undertaken in six countries (Australia, Singapore, Finland, Netherlands, Costa Rica, and the United States). It was expected that differential item functioning (DIF) would exist, and its presence was examined using the ACER Conquest software (Adams, Wu, & Wilson, 2006). The check for DIF involved the examination of the interaction between the group variable *country* and the *item* variable. Then, because of the large number of items, a visual method of presenting item parameter drift (Wu & Adams, 2005) was preferred. Small parameter drifts were identified as statistically different under the hypothesis of zero DIF, but a scatterplot of item difficulties for each pair of countries, as shown in Fig. 8.10, illustrated the stability of the item parameter estimates as an example of differential test function. These procedures are elaborated in the second volume of papers documenting the ATC21S project methodology and research background (Griffin & Care, 2014).

This approach acknowledged that in reality items tend to behave in (at least slightly) different ways for all subgroups, and the majority of items show DIF when the sample is large enough. Consequently, the decision to accept or reject an item based on DIF will still need to be made somewhat subjectively. Those items with item parameters outside the 95% confidence bands in the scatterplot were removed. The Finland data appeared to be affected by language issues when paired with Australia and United States data. Details are provided in the project Volume 2 (Griffin & Care, 2014). In view of the detail and limitations of the item level DIF, it was decided to use differential test functioning (DTF; Badia, Prieto, & Linacre, 2002). The stability of the test function over curriculum, country, and language was encouraging. DIF investigates the items in a test, one at a time, for signs of interactions with sample characteristics. A DIF procedure assessed whether items

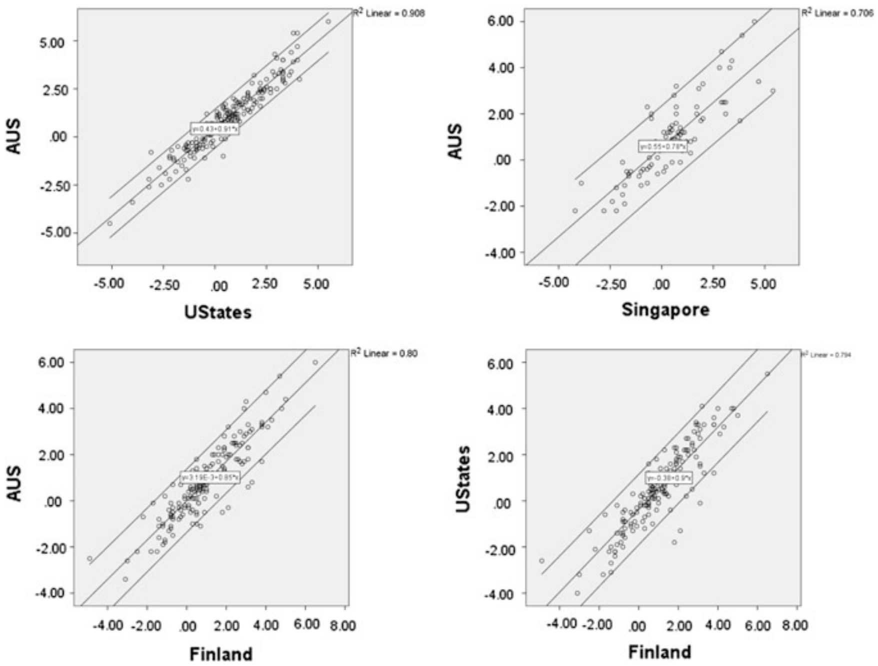


Fig. 8.11 Mapping indicator difficulty across countries

functioned in different ways for different groups. Item functioning is intended to be invariant with respect to irrelevant aspects of the target group, such as, in this case, country, language, and curriculum. But item functioning can be altered by interventions targeted at specific items, such as national curriculum or exposure to the item content. The Mantel-Haenszel procedure (Linacre & Wright, 1989) references two groups at a time to determine whether they differ in a discernible way. A consequence of that analysis is that the effect of accumulated DIF across items for the overall instrument is unclear. DTF (Wright & Stone, 1979) compared the relative difficulty estimates of items obtained from separate analyses, because it provides separate item hierarchies, and the pairwise measures of the group of items are estimated in the context of their own hierarchies. For this reason, the items (categories of data) common across tasks and countries were assessed using DTF. The results of the DTF analysis are illustrated in Fig. 8.11.

The data indicated that the sets of item difficulty estimates used in assessing collaborative problem solving are stable across the six countries. Given that the development sites included three languages, six school curricula, and large and small countries, the stability of the sets of item parameters and their lack of drift was remarkable. It is possible to argue that the ATC21S project constructed a series of collaborative problem-solving tasks delivered via the Internet that measured various aspects of collaborative problem solving as defined by Hesse et al. (2014) in a similar manner across those six countries. Given that the purpose of the ATC21S project was formative (assessment and teaching) and intended to provide teachers with information they could use for instructional purposes, the item parameters were sufficiently stable for this purpose. If the item parameters were also used as a measure of cross-national stability of item difficulty, it might even be possible to use the materials to undertake cross-national comparisons of collaborative problem solving. However, this was not the purpose of the ATC21S project. The project's goal was to establish the psychometric properties of the tasks and to provide teachers with information regarding what a student was ready to learn in developing skills associated with collaborative problem solving.

8.7 Conclusion

There are weaknesses in the initial design that need to be addressed. First the data are restricted to groups of two students working together. In many ways this is a restriction on the kind of task that can be set. If the group size is increased to four or five students, it may be possible to get a measure of internal consistency of the group and to develop measures of individual student performance as well as the group performance.

Despite this, the project has demonstrated that it is possible to separate individual performances within a dyad (group of two) and to identify collaborative problem-solving skills at a reasonably sophisticated level that was sufficiently stable for instructional purposes across country, curriculum, language, and culture. Few measures in education can make such a claim.

The ATC21S project (Griffin et al., 2012) provided some evidence of progress in this field of human-to-human collaboration and problem solving. Large group size will enable within-group variance to be estimated and a separation of the group and individual estimates. Hao and others (this volume) have opened some methodological opportunities. Others have been attempting multilevel item response modeling such as that suggested by Doran, Bates, Bliese, and Dowling (2007) and Salas et al. (this volume) examining the assessment of teams.

Readers of this chapter may feel some frustration with the lack of progress in the modeling and data analytics, but there is some progress and a commitment to proceed. Halpin and von Davier (this volume), for example, examine the application of the Hawkes process to event data collected within dyads. They examine the interpretation in the dyadic context and the appropriateness of an expectation maximization (EM) algorithm for parameter estimation. This is a new research and measurement boundary, and so much remains to be done. The EM algorithm is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a data set when the data has missing values. Such a situation is common in this type of measurement.

References

Adams, R. J., Vista, A., Awwal, N., Scoular, C., & Griffin, P. (2014). Automatic coding procedures for collaborative problem solving. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 115–132). Dordrecht, Netherlands: Springer.

Adams, R. J., Wu, M., & Wilson, M. (2006). *ACER conquest*. ACER: Melbourne.

Badia, X., Prieto, L., & Linacre, J. M. (2002). Differential item and test functioning (DIF & DTF). *Rasch Measurement Transactions, 16*(3), 889.

Baghaei, P. (2007). Local dependency and Rasch measures. *Rasch Measurement Transactions, 21* (3), 1105–1106.

Bentley, T., & Cazaly, C. (2015). *The shared work of learning: Lifting educational achievement through collaboration*. Mitchell Institute Research Report No. 03/2015. Melbourne, Australia: Mitchell Institute.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., et al. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Dordrecht, Netherlands: Springer.

Care, E., Griffin, P., Scoular, C., Awwal, N., & Zoanetti, N. (2014). Collaborative problem solving tasks. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 85–104). Dordrecht, Netherlands: Springer.

Dillenbourg, P., & Traum, D. (2006). Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *Journal of the Learning Sciences, 15*(1), 121–151.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model with the lme4 package. *Journal of Statistical Software, 20*(2), 1–18.

Fischer, A., Greiff, S., & Funke, F. (2012). The process of solving complex problems. *The Journal of Problem Solving, 4*(1), 19–42.

Glaser, R. (1983). *Education and thinking: The role of knowledge*. Technical Report No. PDS-6. Pittsburg, PA: University of Pittsburgh.

Griffin, P. (2014). Performance assessment of higher order thinking. *Journal of Applied Measurement, 15*(1), 53–68.

Griffin, P., & Care, E. (2014). *Assessment and teaching of 21st century skills: Methods and approach*. Dordrecht, Netherlands: Springer.

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 1–15). Dordrecht, Netherlands: Springer.

Griffin, P., & Robertson, P. (2014). Writing assessment rubrics. In P. Griffin (Ed.), *Assessment for teaching* (pp. 125–155). Melbourne, Australia: Cambridge Press.

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2014). A framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 37–56). Dordrecht, Netherlands: Springer.

Kenworthy, L., & Kielstra, P. (2015). *Driving the skills agenda: Preparing students for the future*. Economist Intelligence Unit Report. Retrieved May 27, 2016, from http://www.economistinsights.com/analysis/driving-skills-agenda

Kong, S. C. (2011). An evaluation study of the use of a cognitive tool in a one-to-one classroom for promoting classroom-based dialogic interaction. *Computers & Education, 57*(3), 1851–1864.

Linacre, J. M., & Wright, B. D. (1989). Mantel-Haenszel DIF and PROX are equivalent! *Rasch Measurement Transactions, 3*(2), 51–53.

O'Neil, H. F., Chuang, S., & Chung, G. K. W. K. (2003). Issues in the computer-based assessment of collaborative problem solving. *Assessment in Education: Principles, Policy & Practice, 10,* 361–373.

Organisation for Economic Co-operation and Development. (2013). *PISA 2015: Draft collaborative problem solving framework*. Paris, France: OECD. Retrieved May 27, 2016, from https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Polya, G. (1973). *How to solve it: A new aspect of mathematical method*. Princeton, NJ: Princeton University Press.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press. Reprinted with Foreword and Afterword from *Information and control* by B. D. Wright, 1960. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *The Journal of the Learning Sciences, 14*(2), 201–241.

Verhelst, N. D., & Verstralen, H. H. F. M. (1997). *Modeling sums of binary responses by the partial credit model*. Cito Measurement and Research Department Report No. 97-7. Arneim, Netherlands: Cito.

Vygotsky, L. (1978). *Mind and society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716–730.

Woods, K., Mountain, R., & Griffin, P. (2014). Linking developmental progressions to teaching. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 267–292). Dordrecht, Netherlands: Springer.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA.

Wu, M., & Adams, R. (2005). *Applying the Rasch model to psychosocial measurement: A practical approach*. Melbourne, Australia: Educational Measurement Solutions.

Zoanetti, N., & Griffin, P. (2014). Log-file data as indicators for problem-solving processes. In J. Funke & B. Csapo (Eds.), *The nature of problem solving*. Paris, France: OECD.

Chapter 9
Initial Steps Towards a Standardized Assessment for Collaborative Problem Solving (CPS): Practical Challenges and Strategies

Jiangang Hao, Lei Liu, Alina A. von Davier, and Patrick C. Kyllonen

Abstract Collaborative problem-solving (CPS) skill is an important 21st century skill (Griffin, McGaw, and Care, 2012). However, assessing CPS, particularly in a standardized way, is challenging. The type of collaboration, size of teams, and assessment domain all need to be properly considered when developing a CPS assessment. In this chapter, we outline some practical challenges for developing a large-scale, standardized assessment for CPS and present some strategies to address those challenges. We illustrate these strategies with the Collaborative Science Assessment Prototype (CSAP) developed at Educational Testing Service.

Keywords Collaborative problem-solving · Collaborative science assessment · Tetralogue · Trialogue · Simulation

9.1 Introduction

Collaboration is a "coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem" (Roschelle & Teasley, 1995, p. 70). Compared to individual work, collaboration has several clear advantages: more effective labor division; increased coverage of knowledge, perspectives, and experiences; and enhanced creativity stimulated by the ideas of other group members (Organization for Economic Co-operation and

This work was conducted while Alina A. von Davier was employed with Educational Testing Service.

J. Hao (✉) · L. Liu · P.C. Kyllonen
Educational Testing Service, Princeton, NJ 08541, USA
e-mail: jhao@ets.org

A.A. von Davier
ACT, Iowa City, IA 52243, USA
e-mail: Alina.vondavier@act.org

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_9

Development, 2013). Collaborative problem solving (CPS) is a special type of collaboration. In educational settings, it may be defined as a process that includes both cognitive and social practices in which two or more peers interact with each other to share and negotiate ideas and prior experiences, jointly regulate and coordinate behaviors and learning activities, and apply social strategies to sustain the interpersonal exchanges to solve a shared problem. This definition describes CPS as both a cognitive and social process (Dillenbourg, Järvelä, & Fischer, 2009; Järvelä, Volet, & Järvenoja, 2010; Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015; Van den Bossche, Gijselaers, Segers, & Kirschner, 2006).

Despite its advantages, collaboration does not necessarily lead to better results or improved productivity even for a team of capable individuals. It is important that team members collaborate effectively. Even for a set of capable individuals, there are successful collaborations that lead to improved results or increased productivity, and unsuccessful ones that lead to worse results or decreased productivity. It is therefore plausible to assume that there is a certain CPS skill that can lead to a successful collaboration. This CPS skill "is the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution" (OECD, 2013, p. 6).

CPS may be considered a critical skill for academic and career success in the 21st century (Griffin, McGaw, & Care, 2012). The majority of studies on CPS have focused on learning, for example, finding effective ways to promote learning in a (computerized) collaborative environment (Koschmann, 1996; Stahl, Koschmann, & Suthers, 2006) or developing tasks to foster collaboration skills that improve learning (Sottilare, Brawner, Goldberg, & Holden, 2012). In contrast, the assessment aspect of CPS has been relatively less covered. The precise measurement of CPS skill is challenging and includes addressing psychometric requirements for assessments, such as validity, reliability, and fairness.

Among the existing studies on assessing CPS, most of them are designed from the perspective of revealing important aspects of CPS (Cohen, Lotan, Scarloss, & Arellano, 1999; DeChurch & Mesmer-Magnus, 2010; O'Neil, 2014; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010). A recent review of studies along this line can be found in von Davier and Halpin (2013). Studies often do not use standardized assessments, that is, ones in which items, scoring procedures, and interpretations are consistent across test forms, and test administrations themselves are predetermined and standardized. However, one exception is the standardized CPS assessment developed for the Programme for International Student Assessment (PISA) in its sixth survey during 2015 (OECD, 2013). In this assessment, students collaborated with a different number of virtual partners (avatars) on a set of computer-based collaborative tasks and communicated with their virtual partners by choosing from a list of predefined texts. The use of virtual avatars and predefined texts is a compromise from a person-to-person collaboration made to ensure standardization. Another notable assessment for CPS (not standardized) was developed for the Assessment and Teaching of 21st Century Skills project (ATC21S) carried out by Griffin and colleagues (Care & Griffin, 2014; Griffin et al., 2012). In this assessment, two students collaborated via text chat to solve computer-based collaborative tasks. Their chat communications, keystrokes, and response time were automatically coded according to a CPS framework (Adams et al., 2015). The final CPS assessment results from both PISA 2015 and ATC21S have not yet been published at the writing of this book chapter.

Developing a standardized assessment for CPS is extremely challenging. The goal of this chapter is to discuss the challenges in designing standardized collaborative assessments and to propose several strategies to mitigate these challenges. We illustrate our proposed strategies with a particular assessment prototype, the Collaborative Science Assessment Prototype (CSAP) (Hao, Liu, von Davier, & Kyllonen, 2015), which was developed at Educational Testing Service (ETS) to assess collaborative problem-solving skills in the domain of science. We do not present comprehensive findings from the project here, but focus instead on illustrating the implementation of the proposed strategies.

9.2 Practical Challenges and Proposed Strategies

It is challenging to assess CPS, particularly as a standardized assessment. An assessment is essentially an instrument used to measure certain predefined constructs using the evidence exhibited by the test takers during their interaction with the assessment components. To produce appropriate and useful types of evidence, the assessment components need to be designed carefully, for example, by following recommendations from an evidence-centered design (ECD) framework (Mislevy & Riconscente, 2006).

9.2.1 CPS Construct Definition

The first practical challenge for assessing CPS is to define clearly the complex CPS constructs. CPS involves various facets, some cognitive and some noncognitive or social. What makes the situation more complicated is that each facet may exhibit itself differently in different tasks, domains, and team compositions. A strategy for addressing this challenge is to define clearly which specific facets of CPS will be measured in a given assessment. For example, PISA 2015 considered three critical CPS skills, establishing and maintaining shared understanding, taking appropriate action to solve the problem, and establishing and maintaining team organization (OECD, 2013). The ATC21S targeted five skills: participation, perspective taking, social and task regulation, and knowledge building (Hesse, Care, Buder, Sassenberg, & Griffin, 2015). The assessments developed for both PISA 2015 and ATC21S were designed to assess broad, domain-general skills. In contrast, in our work we built a domain-specific assessment focusing on four CPS skills for collaboration around tasks within the science domain (Liu et al., 2015) as it will be described later in this chapter.

9.2.2 Complex Relationship Between Evidence and CPS Constructs

The second practical challenge is the complex relationship between the evidence from specific tasks for assessing CPS constructs. Generally speaking, in a collaborative task, there are two types of directly observable evidence: the test takers' responses to the task and the communications among test takers during the collaboration. Here, the responses can be the choices as in multiple-choice items or text input as in constructed-response items. They can also be the time-stamped processes of the test takers' responses to a complex/interactive item, such as a game or a simulation. The communication among team members can occur through different modes, for example, text chat, audio or video, or face-to-face. The inference about the underlying constructs of CPS must be based on directly-observable evidence (As determined through the ECD process).

For individual-based assessments, a given test taker's performance depends solely on the properties of the items, test taker's responses to the items, and some objective conditions of measurement. In this case, when the test is well-designed, the mapping between the evidence and construct is relatively straightforward, though not necessarily simple. In contrast, for an assessment in a collaborative setting, the issue becomes much more complex. To illustrate this complexity, let us consider the simplest collaborative case, the dyadic collaboration. Let us denote the two participants in a dyadic team as A and B. To directly record the effect of collaboration, we also assume we capture their responses to items both before and after collaboration. Even in such a simple setup, the possible mapping between the observable evidence and constructs can be complex because the responses from one person depend both on the person's individual skills and on the interaction with the other team member. The interactions between the team members are subject to both team members' cognitive skills and collaborative skills. As a result, the response to an item after collaboration will depend on the properties of the item, the test taker's own cognitive skill, the teammate's cognitive skill, and the CPS skills of both team members, in addition to some non-CPS personality factors and conditions of measurement. It will also depend on the implementation of a specific collaborative process. Figure 9.1 summarizes the possible relationships among constructs and evidence. Note that this illustration reflects only the simplest collaboration, the dyadic collaboration. When there are more members, the complexity of the dependencies will increase significantly.

In Fig. 9.1, we only illustrate the possible dependencies among the evidence and the constructs. In practice, these mappings can also be domain-dependent and task-dependent, raising empirical questions that need to be addressed based on data from particular tasks and domains. All these dependencies make the inference of constructs from evidence very complex and challenging.

A possible, albeit limited, strategy to address this challenge is to break the complex interdependencies through the assessment design: either through a separate assessment for each individual's cognitive skill in the same domain as the CPS

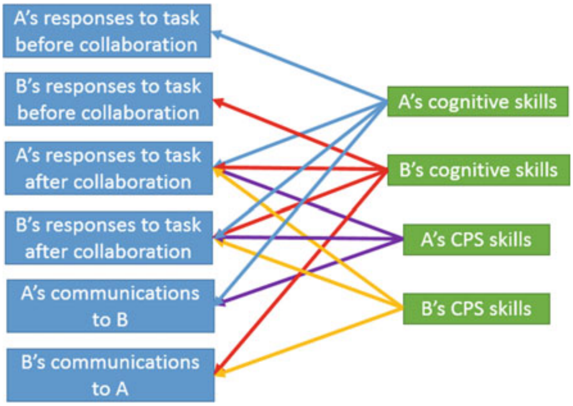

A's responses to task before collaboration
B's responses to task before collaboration
A's responses to task after collaboration
B's responses to task after collaboration
A's communications to B
B's communications to A
A's cognitive skills
B's cognitive skills
A's CPS skills
B's CPS skills

Fig. 9.1 The possible relationship between the evidence and constructs

assessment targets, or by providing opportunities for each member of the team to respond to the items before collaborating, or both. In our development of the CSAP, we chose to do both. However, the CSAP did not allow for the in-depth investigation of the dependencies due to task's specifics or to the domain. These will be investigated in a future study.

9.2.3 Credit Assignment

The third practical challenge is how to assign credit fairly to each team member based on collaborative work. A major goal of assessment is to provide scores that reflect the proficiency on the targeted construct at the individual or group level. For a CPS assessment, ideally, we would report certain CPS scores that reflect the CPS skills for each individual test taker. However, each individual's exhibition of CPS skills will be affected by the other members in the team. More importantly, teamwork may be seriously affected if there are uncooperative team members, making it a significant challenge to establish a fair way to assign credit in collaborative work.

One possible strategy to providing individual scores is to place an individual into several different teams with different, carefully sampled partners. Then, based on the performance of all these collaborations, one could map out a distribution of CPS skills of the individual test taker. Obviously, such an arrangement requires significant effort to balance the test design so as to ensure that a sufficient number of people are present during one administration of the test so that the test can include multiple tasks and multiple partners for team work. Clearly, this type of design is extremely difficult in practice, at least at the moment. One alternative strategy is to use virtual intelligent agents of varied characteristics, as was done in PISA 2015. However, currently available artificial intelligence technology is still not able to support realistic open conversations with real humans. A possible compromise may be some clever combination of human-human collaborations and human-agent collaborations. In our CSAP we consider only human-human collaboration as a start.

On the other hand, if we step back a bit from reporting the CPS skills of individuals and focus on the statistical properties of the CPS skills of many teams, we may simplify the situation considerably. For example, the difference in overall CPS skills of people from different schools or companies may also be of interest. In this case, the simplest design will be to randomly sample students from the same large group (e.g., school or company) and assign them into teams to complete the CPS tasks. The CPS performance of the large group will be a certain aggregation of the CPS skills exhibited by each team. As long as there are a sufficient number of teams in each large group and members are randomly selected, the group level statistics could be representative, although to our knowledge there has yet not been an evaluation of the representativeness of a distribution of skills generated by random pairings of group members. Nevertheless, in this case, one might not need to worry about how to assign credit fairly to each individual, as one would be interested in the aggregated results. The random assignment should tend to balance out different effects. Such an arrangement may find a realistic use in the real world. For example, big companies may be interested in finding out the average CPS skills of their employees rather than of individuals to inform their workforce development. In our CSAP, we mainly focus on the statistical properties of CPS skills from teams rather than individuals.

9.2.4 Confounding Factors

The fourth practical challenge is the existence of potential confounding factors, such as gender, culture background, and language proficiency, which could affect both the process and outcome of collaboration (Kreijns, Kirschner, & Jochems, 2002; Sycara, Gelfand, & Abbe, 2013; Van den Bossche et al., 2006). A possible strategy to address this challenge is to choose an appropriate channel for communication to reduce these confounding factors during the collaboration. Among the various common ways of communication (such as audio, video, face-to-face, and text chat), text chat is probably least likely to reveal personal biometric and background information, and therefore, may be best in mitigating the potential effects of confounding factors. However, to what extent text-chat-mediated communications can approximate face-to-face communications remains an open question and should be addressed by empirical studies. At the moment, another clear advantage of text chat communication is its technological feasibility, in both communication bandwidth and potential for automated processing and scoring. In addition, privacy concerns are less prominent in text-mediated communication as compared to other communication means (e.g., video and audio). In our CSAP, we choose to use text-mediated communication.

9.2.5 Effects of Team Composition

The fifth practical challenge concerns team composition. The number of members of a team and their cognitive and CPS skills can all potentially affect the collaboration process and thus affect the exhibition of the CPS skills of other members in the team. It is generally believed that working with partners who have different levels of cognitive skills relevant to the task will improve collaboration outcomes (Webb, Nemer, Chizhik, & Sugrue, 1998). The balance in team composition, with respect to cognitive and CPS skills likely will affect the results of the assessment. The number of team members will also affect the collaboration outcomes and process. Increasing the number of team members has been shown to lead to increased social loafing (Karau & Williams, 1993). A possible strategy to addressing this challenge may be to assign team members with balanced domain-relevant cognitive skills levels or some other meaningful factors (such as personality), and also limit the number of team members to small numbers, say, two or three. (Team size may be an important research topic in itself.)

9.2.6 Selection of Tasks

The last, but not least, practical challenge we consider in this chapter is the selection of tasks or items in the CPS assessment. To measure CPS, we must have a set of tasks or items that allow people to solve collaboratively. The development and selection of tasks or items is crucial. In collaboration, students need to work together to establish a common stage of reference, identifying discrepancies in understanding, negotiating to resolve those discrepancies, and developing a joint understanding (Barron, 2003; Roschelle, 1992). A common view is that the tasks or items for collaboration should be *group-worthy*, by which is meant the following (Lotan, 2003):

- They are open-ended and require complex problem solving.
- They provide students with multiple entry points to the task and multiple opportunities to show intellectual competence.
- They deal with discipline-based, intellectually important content.
- They require positive interdependence as well as individual accountability.
- They include clear criteria for the evaluation of the group's product.

When we select tasks, we must account for the distinction of collaboration for assessment versus collaboration for learning. There is little doubt that group-worthy tasks will make collaborative learning more effective. From the assessment perspective, group-worthy tasks may create situations in which the collaboration is *scaffolded* or *forced*. *Scaffolding* refers to the provision of partial information to assist learners. Scaffolding has been shown in learning to be beneficial, but it is not always useful in assessment. Overscaffolding in a CPS assessment, that is providing extensive aid to support collaboration, may teach people how to collaborate, and therefore, may lead to overestimated CPS skills. Moreover, the open-ended nature of group-worthy tasks adds additional complexity to having equal opportunity for each team member to exhibiting the CPS skills. *A forced* collaboration refers to one in which participants are forced into collaboration. For example, one of the typical group-worthy tasks is the so-called jigsaw task, in which each team member will get part of the information needed to complete the task. In this situation participants must collaborate so as to have sufficient information to complete the task. However, from an assessment perspective, one needs to design the task in a careful way to ensure that different partial information obtained by each team member will not disadvantage the participant in the scoring of the collaboration. It is a challenge for the task designer and requires more empirical iterations to validate the task. ECD can be extended here to incorporate the task model, student model and team model, for example, as shown in Kerr, Andrews, and Mislevy (in press).

To address this challenge, one needs to find the sweet spot between group-worthy and over scaffolding, while keeping participants' roles in the task as balanced as possible. One practical strategy may be sacrificing some group-worthy features for balanced roles and providing a controllable system instruction or facilitation to ensure that team members are engaged in the collaboration. Again, determining the appropriate level of facilitation is an empirical endeavor, and we need to carry out actual studies to find the optimal level of facilitation. Including a facilitator (agent) or a system-prompt is the strategy adopted in our prototype

So far, we have outlined several practical challenges for developing an assessment for CPS and provided our considerations on the possible strategies for mitigating these challenges. We need to emphasize that this list of challenges is far from complete and the strategies we propose are not necessarily optimal. However, not all the issues can be solved by purely theorizing, and actual empirical research is needed to expose the challenges and find better strategies.

9.3 Collaborative Science Assessment Prototype (CSAP)

In the previous sections, we outlined various challenges for developing a standardized CPS assessment, proposed several strategies to address these challenges, and briefly referred to the work we conducted on the CSAP. In this section, we describe this collaborative science assessment prototype in more detail and we show how this CSAP embodies the envisioned strategies in order to investigate various aspects of the CPS assessment. We emphasize that this CSAP is a prototype, which allowed us to get a better sense of the challenges and explore the feasibility of our strategies in practice.

9.3.1 Assessment Instruments

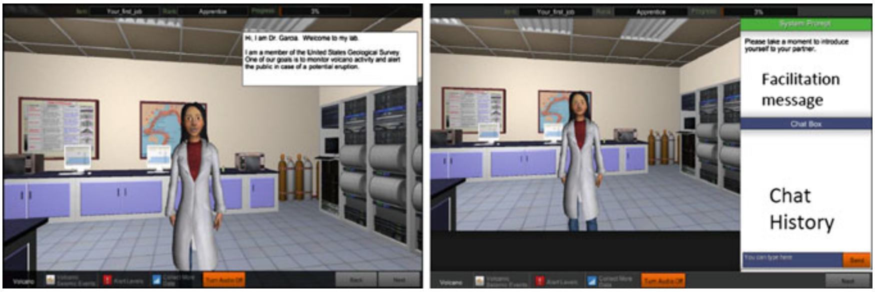
The CSAP project was designed to measure CPS skills in the science domain, addressing the six practical challenges following the strategies we introduced earlier. We introduce the assessment instruments used in this study and then show how the six strategies are implemented. Five assessment instruments were administered in the CSAP study[1]:

- A standalone test for general science knowledge consisting of 37 multiple-choice items adapted from the Scientific Literacy Measurement (SLiM) instrument (Rundgren, Rundgren, Tseng, Lin, & Chang, 2012).
- A personality survey, Ten Item Personality Measure (TIPI) (Gosling, Rentfrow, & Swann, 2003).
- A demographic survey adapted from the National Assessment of Educational Progress (NAEP, 2013).
- Two versions of a web-based science simulation task on volcanoes.
- Collaborative version (a.k.a. Tetralogue): Two participants collaborate to interact with two virtual agents in the simulation to complete a science task on volcanoes.
- Single-user version (a.k.a. Trialogue): A single participant interacts with two virtual agents in the simulation to complete a science task on volcanoes.
- A postcollaboration satisfaction survey.

The two simulation tasks were both modified from an existing simulation, the *Trialogue* (Zapata-Rivera et al., 2014). This simulation was designed based on ECD (Mislevy & Riconscente, 2006) to measure students' scientific inquiry skills using multiple-choice (MC), constructed-response (CR), and conversational items. In the Trialogue simulation, a student interacts with two virtual agents (one serves as a student peer and another serves as a mentor) to complete a set of (science) tasks about volcanos. The name *Trialogue* describes the conversations between the student and two virtual agents (Feng, Stewart, Clewley, & Graesser, 2015). The single-user version was used for two purposes: (a) it served as a control to check the effect of collaboration, and (b) we used the responses in the single-user version to provide a baseline for item properties, such as the item proportion correct. The collaborative version of the simulation, the Tetralogue, included a chat-window to allow two test takers to communicate with each other, in addition to the chat window that allows the team to communicate with the agents; hence, the name the Tetralogue refers to the four in this simulation.

In both versions of the simulations, the time-stamped responses to the questions and all turn-by-turn communications were recorded into a carefully designed log file (Hao, Smith, Mislevy, von Davier, & Bauer, 2016a). The conversations were used to measure the CPS skills and the responses to the in-simulation items were

[1] We introduced all the components in this study in this chapter, but the findings from several of them (e.g., personality, general science-knowledge test) won't be reported here.



Hi, I am Dr. Garcia. Welcome to my lab.

I am a member of the United States Geological Survey. One of our goals is to monitor volcano activity and alert the public in case of a potential eruption.

Please take a moment to introduce yourself to your partner.

Fig. 9.2 Screenshots from the collaborative simulation task

used to measure science inquiry skills (Zapata-Rivera et al., 2014). Note that the interactions between human and virtual avatars are not included in our analysis, mainly because we focused on the human-human interactions first. In Fig. 9.2, we show screenshots from the single-user (left) and the collaborative (right) versions of the simulation task. The major difference between the single-user and collaborative version is the additional chat box for communication in the collaborative version.

9.3.2 Implementing the Strategies

9.3.2.1 CPS Construct Definition

To apply the six strategies to address the aforementioned challenges, first, we focus on the CPS skills in the domain of science. Though both PISA 2015 and ATC21S consider CPS skills to be less domain dependent, we took a strategy of measuring the CPS skills within a specific domain, that of science. Based on the findings from the research on computer-supported collaborative learning (CSCL; Barron, 2003; Dillenbourg & Traum, 2006), the Collaborative Problem Solving Framework from PISA 2015 (OECD, 2013), and ATC21S (Griffin et al., 2012), we developed a CPS framework by targeting the CPS skills in the domain of science (Liu et al., 2015). There are four CPS skills being targeted in our study: sharing ideas, negotiating ideas, regulating problem-solving activities, and maintaining communication. Each of these major categories has subcategories, yielding a total of 33 subcategories. A summary of the coding rubrics can be found in Table 9.1.

9.3.2.2 Complex Relationship Between Evidence and CPS Constructs

To disentangle the complex interdependencies among team members' CPS skills and cognitive skills, we used the general science-knowledge test to provide a separate assessment for each individual's general science skill. Moreover, we designed a four-step response procedure in the collaborative version of the

Table 9.1 Coding rubric of CPS skills used in this paper

CPS skills	Student performance (subcategories)
Sharing ideas	1. Student gives task-relevant information (e.g., individual response) to the teammates
	2. Student points out a resource to retrieve task-relevant information
	3. Student responds to the teammate's request for task-relevant information
Negotiating ideas	4. Student expresses agreement with the teammates
	5. Student expresses disagreement with teammates
	6. Student expresses uncertainty of agree or disagree
	7. Student asks the teammate to repeat a statement
	8. Student asks the teammate to clarify a statement
	9. Student rephrases/complete the teammate's statement
	10. Student identifies a conflict in his or her own idea and the teammate's idea
	11. Student uses relevant evidence to point out some gap in the teammate's statement
	12. Student elaborates on his or her own statement
	13. Student changes his or her own idea after listening to the teammate's reasoning
Regulating problem solving	14. Student identify the goal of the conversation
	15. Student suggests the next step for the group to take
	16. Student expresses confusion/frustration or lack of understanding
	17. Student expresses progress in understanding
	18. Student reflects on what the group did
	19. Student expresses what is missing in the teamwork to solve the problem
	20. Student checks on understanding
	21. Student evaluates whether certain group contribution is useful or not for the problem solving
	22. Student shows satisfaction with the group performance
	23. Student points out some gap in a group decision
	24. Student identifies a problem in problem solving
Maintaining communication	25. Student responds to the teammate's question (using texts and text symbols)
	26. Student manages to make the conversation alive (using texts and text symbols, using socially appropriate language)
	27. Student waits for the teammate to finish his/her statement before taking turns
	28. Student uses socially appropriate language (e.g., greeting)
	29. Student offers help
	30. Student apologizes for unintentional interruption
	31. Student rejects the teammate's suggestions without an accountable reason
	32. Student inputs something that does not make sense
	33. Student shows understanding of the teammate's frustration.

simulation task, by which we can capture each team member's science inquiry skills before and after the collaboration. The procedure for responding to a question in the simulation is as follows:

1. Each participant is prompted to respond to the item individually before any collaboration.
2. Each participant is prompted to discuss the item with her partner.
3. Each participant is prompted to revise her initial response if she wants.
4. A representative is randomly chosen to submit a team answer.

In this way, the responses before collaboration capture each individual member's science inquiry skills specific to the task, while the changes in responses after the collaboration reflect how effective the collaboration was and allow us to probe directly which CPS subskills may be more important for better collaboration outcomes. It is worth noting that the fourth step in the above procedure is mainly designed to avoid deadlocks in the collaborations. In case an agreement cannot be achieved, we needed a mechanism to move on to the next item.

9.3.2.3 Credit Assignment

To address this challenge, in some of our analyses we focused on the team as the unit of measurement and present the statistical properties of the CPS skills from many teams. As illustrated in previous section, for this purpose, we do not need to place each test taker into many teams. We randomly assign participants to dyadic teams. This is a compromise we made due to the constraints imposed by time, budget and technological infrastructure we had for this first prototype and initial exploratory study. The findings also provide information for our future endeavor to measure individual level CPS skills. There are other attempts to look at the data from the individual process data to investigate the engagement (Halpin, von Davier, Hao, & Liu, in press) and to investigate the propensity for each person to display a particular collaborative behavior (Andrews et al., in press). These studies are not discussed here.

9.3.2.4 Confounding Factors

To address this challenge, we chose the text chat as our means of communication and also administered a personality and demographic survey to each participant to measure the factors such as cultural background, gender, and personality.

9.3.2.5 Team Composition

To address this challenge, we chose to start with the simplest setup (i.e., limiting the number of team members to two and randomly assigning the team members to each dyad).

9.3.2.6 Selection of Tasks

To address this challenge, we chose one simulation-based task that included a simpler form of collaboration and was developed carefully following the ECD process. We also hypothesized that educational simulations that provide complex digital environments are more likely to elicit collaborative work. However, this particular task is only minimally group-worthy as it is modified from an existing simulation, initially designed for a single user who interacts with two virtual avatars in the simulation (Zapata-Rivera et al., 2014). To ensure that the participants engage in collaboration, we designed a set of facilitation messages to prompt the team members to collaborate. Specifically, the aforementioned four-step response procedure was developed to facilitate the flow of the collaboration.

9.3.3 Data Collection and Scoring

We collected the data through Amazon Mechanical Turk, a crowdsourcing data collection platform (Kittur, Chi, & Suh, 2008). We recruited 1500 participants located in United States with at least one year of college education. We administered to them the general science test, personality survey, and demographic survey. Then we randomly selected 500 to take the single-user version of the simulation. The remaining 1000 were randomly paired into dyads to take the collaborative version of the simulation. The data from the simulation task for each team include both the responses to the items in the simulation and the text chat communication between the team members around each item. There are seven multiple-choice-like items in the simulation task, and for each item, there are about five turns of conversation. After removing incomplete responses, we had data from 483 dyads. The responses to the seven multiple-choice-like items were scored based on the corresponding scoring rubrics as presented in Zapata-Rivera et al., (2014). In addition to scoring the outcome responses, we also annotated the chat communication during the collaboration based on our CPS framework (Liu et al., 2015). Two human raters were trained on the CPS framework, and they double-coded a subset of discourse data (15% of the data). The unit of analysis was each turn of a conversation, or each conversational utterance. The raters had two training sessions before they started independent coding. In the first session, the raters were trained on the 33 subcategories of the CPS framework using the skills definitions and coding examples for each subcategory. In the second training session, the trainer and two raters coded data from one dyad together to practice the application of specific codes and address issues specific to classifying utterances using the CPS framework. After the training sessions, the two raters independently coded the discourse data from 79 dyads. One of the 33 subcategories was assigned for each turn, and the inter-rater agreement in terms of unweighted kappa was 0.61 for all 33 subcategories. Based on the subcategories, we derived the corresponding four major categories of the CPS skills based on Table 9.1. The inter-rater reliability in terms of unweighted kappa based on the four major categories was 0.65. According to Fleiss and Cohen (1973), a kappa value of 0.4 is an acceptable level of agreement for social science experiments. Currently, efforts are being made to develop an automatic scoring engine for these chats (Flor, Yoon, Hao, Liu, & von Davier, 2016). The engine will be trained on the data and on the scores from the human scorers. The remaining data were coded by one rater.

Among the 483 dyads' responses, a further scrutiny of the data showed that many of the teams started some conversations even before the system prompted them to discuss. This means that they started conversations before or during the period that they are supposed to make initial responses individually. Different teams had nonprompted conversations for a different subset of the items, which complicates the analysis. Of the teams, 82 did not have any nonprompted conversations while the rest had at least some non-prompted discussions for a varying number of items. We compared the scores of the general science knowledge test for participants from the 82 teams with the scores for the rest of the teams via a two-tailed t-test for independent samples and found that the two groups were not different ($p = 0.38$). However, we focus here on the data from the 82 teams.

9.3.4 Quantifying the Collaboration Outcomes

The difference between the revised response and initial responses is a measure of collaboration outcomes. If we treat each dyad as the unit of analysis, we need to define variables to quantify the answer changes for each item. We first introduce the *number of changes* (denoted as n) to quantify how many revised responses are different from initial responses from both members of each dyad for each item. The possible values for n are {0, 1, 2}: $n = 0$ when nobody makes any changes, $n = 1$ when only one person makes changes, and $n = 2$ when both members make changes. Next, we introduce *score change* (denoted as s) to quantify the total score changes between the revised response and the initial response from both members of each dyad for each item. The definition of s is the sum of the score difference between initial responses and revised responses for the two members of each dyad, that is $s = (r_1 - i_1) + (r_2 - i_2)$, where r_j and i_j denote a revised and an initial response to an item, respectively, for person j. The possible values for s are {−2, −1, 0, 1, 2}. One should note that for the state $s = 0$, there are two different possibilities. The first is that both members do not change their responses. The second is that one member changes a response from incorrect to correct and the other changes from correct to incorrect. Therefore, to have a complete description of the changes at a dyadic level, we introduce the vector "item collaboration effect" for each item, $\delta_k = (s_k, n_k)$, with δ_k defined at the item level and subscript k denoting the item number. At the task level, we simply sum all items, which gives $\Delta = (S, N)$, where $S = \sum_k s_k$ and $N = \sum_k n_k$. By convention, we use the lowercase *n* and *s* to denote the item level changes and the uppercase *N* and *S* to denote the task-level changes.

It is worth noting that the change of responses after collaboration does not necessarily mean the cognitive skill of each member of the team improved. The change only represents the effect due to the interaction of the team members. That is, better communication (collaboration) among the members may improve the response as people learn (hear) from each other and reflect on their initial thoughts. However, if you dismiss the team and ask members to take a parallel test individually again, each member is likely still at his or her previous level, as a short period of collaborative work is not likely to change the cognitive skill. Whether these changes are consistently generalizable across different tasks is an empirical question that needs to be addressed with more empirical data.

9.3.5 Quantifying the CPS Skills

Each turn-by-turn conversation was classified as one of the four categories of CPS skills (e.g., share ideas, negotiate ideas, regulate problem solving, and maintain communication). We introduced a *CPS profile* as a quantitative representation of the CPS skills of each dyad. The profile was defined by the frequency counts of each of the four CPS-skill categories or their combinations and had two levels, unigram and bigram. The unigram, bigram, or even ngram levels are used in natural language processing to represent pieces of text (often, words) that appear alone, two together as consecutive pairs, or more together as a sequence of more pieces of text, respectively. We borrowed this idea to represent the CPS subskills and the sequence in which they are displayed, but we use only unigrams and bigrams as the frequency count is too low for ngrams. The frequency counts of the different CPS subskills were used at the unigram level, while the frequency counts of consecutive pairs of CPS skills in the conversations were used at the bigram level. As such, each dyadic team's communications can be represented by the corresponding CPS profile.

It is worth noting that though we considered only unigrams and bigrams of CPS skills, other collaboration-related information can also be appended to the profile, such as the number of turns and the total number of words. Such a profile is essentially a vector representation of collaboration skills exhibited by each team. The vector nature of this representation allows us to easily calculate similarity or dissimilarity among the teams.

A remaining question is how comparable these ngram frequencies are across different tasks. It maybe plausible to assume that the absolute value of the frequencies are likely to be comparable for similar items, not less comparable for drastically different items. However, the relative frequencies of different ngram features may be more robust across tasks. This is essentially an empirical question that can only be addressed with large amount of empirical data involving different tasks.

9.4 Preliminary Findings

As mentioned, the goal of this chapter is to outline the practical challenges for developing a standardized assessment for CPS and some strategies to address these challenges. A comprehensive report of findings from the CSAP is beyond the scope of this chapter but we present two interesting findings. The first is about the relationship between the property of the item in the simulation task and the amount of collaboration in terms of total number of words and turns in the communication it elicits. The second is about the relationship between the CPS profile and collaboration outcomes (Hao, Liu, von Davier, Kyllonen, & Kitchen, 2016b).

One of the most important pragmatic questions for assessing CPS is to determine what tasks should be used. One may expect that different tasks will elicit a different amount of communication and lead to different collaboration outcomes. For example, it seems plausible that very simple tasks may not elicit much collaboration. Therefore, a proper understanding of the relations between task properties and the amount of collaboration it induces is important for selecting appropriate items for the assessment. In the simulation task used in the CSAP, the first seven items are selective-response items and have binary correct/incorrect scores. We focus on these items in our analysis. As mentioned in our study design, we included a single-user version of the simulation-based task that was completed by individual participants. Their responses to the task allowed us to calibrate the item proportion correct. Based on these individual responses to the first seven items, we obtained a Cronbach's alpha of 0.65.

In Fig. 9.3, we show the results of item proportion correct, average word count, and average number of turns. The results suggest a linear relationship between the item proportion correct and the total number of words and turns in the communication. This relationship provides an informative guideline for choosing appropriate items for CPS task in a CPS assessment.

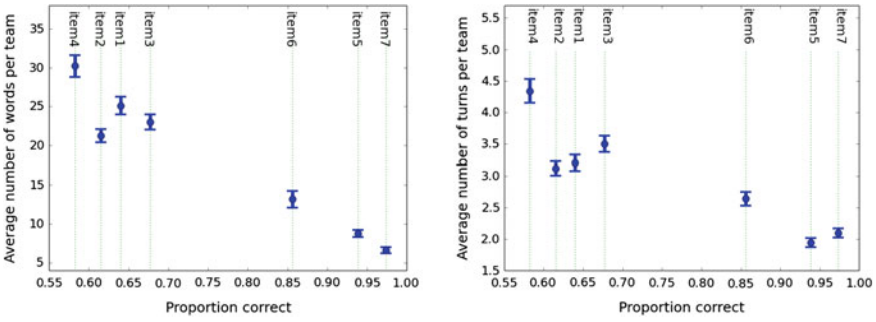
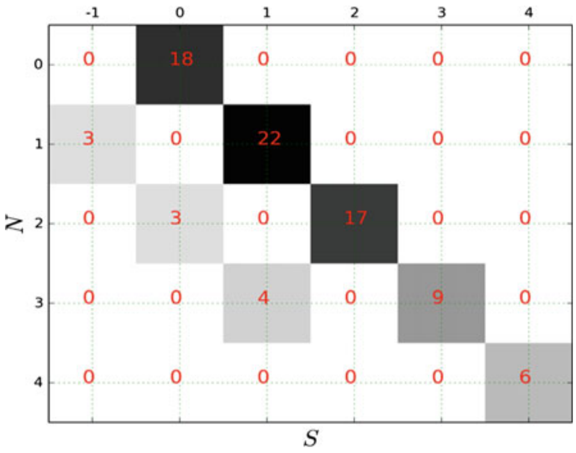


Fig. 9.3 The relationship between item proportion correct and the average number of words (*left*) and an average number of turn-takings (*right*) in the communication. The dots and error bars in the plots are the means and standard error of the means



N \ S	-1	0	1	2	3	4
0	0	18	0	0	0	0
1	3	0	22	0	0	0
2	0	3	0	17	0	0
3	0	0	4	0	9	0
4	0	0	0	0	0	6

Fig. 9.4 The distribution of the teams in space spanned by N and S

To show the relationship between CPS profile and collaboration outcomes, we introduce *effective collaboration* and *ineffective collaboration* based on the variables N and S, to quantify the collaboration outcomes:

- Effective collaboration: $N > 0 \cap S > 0$. (That is, collaboration leads to positive changes.)
- Ineffective collaboration: $(N > 0 \cap S \leq 0) \cup N = 0$. (That is, collaboration leads to negative changes or to no changes.)

Note that the criterion for effective collaboration is not necessarily fixed. In the current study, we considered the collaboration effective as long as at least one member made at least a total net change from incorrect to correct. If nobody in the team made at least one total net correct change, we classified the collaboration as ineffective. Figure 9.4 shows how the 82 teams were distributed in the space spanned by S and N.

Next, we compared mean CPS profiles of the teams from the effective and ineffective collaborations; results are shown in Fig. 9.5. From these results, one can readily see that at the unigram level, the teams with effective collaboration showed significantly more negotiating skills than the teams with ineffective collaboration. At the bigram level, teams with effective collaboration exhibited significantly more of the following consecutive CPS skill pairs: share-negotiate, negotiate-share, regulate-share, and negotiate-negotiate. However, the teams with ineffective collaboration showed many more share-share skill pairs.

9.4.1 Relative Importance of CPS Skills

Figure 9.5 shows that for some CPS skills effective collaborations exhibited more of that skill than ineffective collaborations did, for some CPS skills the opposite was

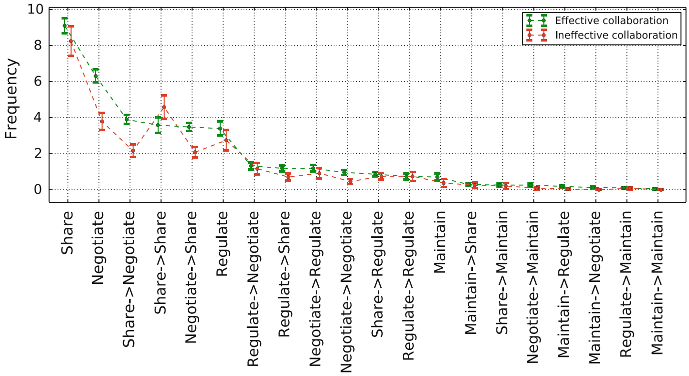


Fig. 9.5 Unigram and bigram profile of CPS skills for the teams corresponding to effective and ineffective collaborations

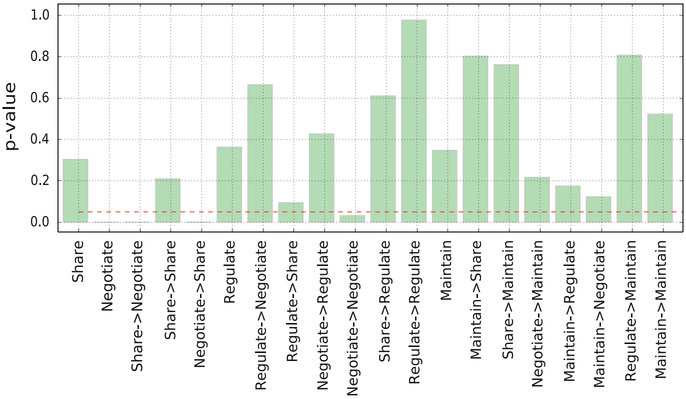


Fig. 9.6 *P*-value of t-test on the frequency of different CPS skills corresponding to effective and ineffective collaborations. The red horizontal dashed line corresponds to a significant level of 0.05

true, and for some skills there was no difference. To get a quantitative measure of the relative importance of each CPS skills (or skill pairs), we used two methods as follows.

First, we performed a t-test for each of the CPS skills (or skill pairs) for the effective collaboration and ineffective collaboration groups. We used the corresponding *p*-value to tell which skills or skill pairs were more discrepant. The *p*-value for each component of the CPS profile was shown in Fig. 9.6. If we choose 0.05 as the significance level, negotiate, share-negotiate, negotiate-share and negotiate-negotiate stand out immediately.

A second method we used was to get the relative importance from a random forest classifier (Breiman, 2001; Ho, 1995) applied to the features variables (e.g., the CPS skills or skill pairs that constitute the CPS profile) with labels corresponding to

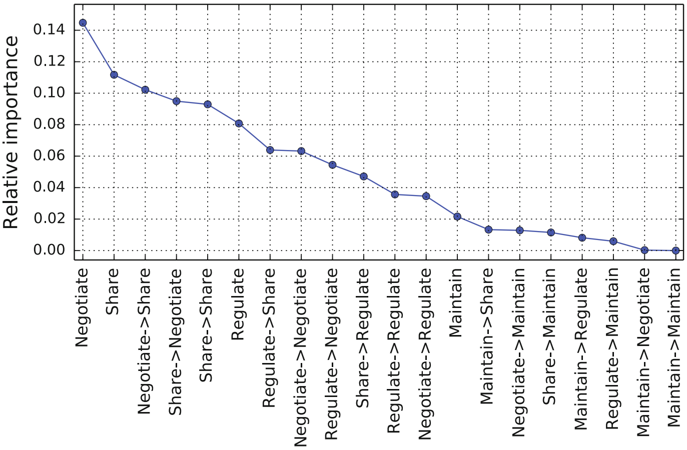


Fig. 9.7 Relative feature importance based on a random forest classifier

effective and ineffective collaborations. During the training of the classifier, a set of decision cuts were made on each feature variable. The relative depth of a feature used as a decision node in a decision tree represents the relative importance of that feature with respect to the predictability of the target labels. Generally speaking, features used at the top level of the decision tree will affect a larger fraction of the sample in terms of the final prediction. Therefore, the expected fraction over the trees in the forest can be used as an estimate of the relative importance of the features. Figure 9.7 shows the relative importance of the CPS skills and skill pairs based on such an analysis. Again, negotiation-related skills top the ranking.

The results from these two different analyses converge nicely on the message that negotiation is a critical skill for a successful collaboration. This result is consistent with findings in the literature on knowledge-building discourse (Scardamalia & Bereiter, 1994; Stahl, 2006), as knowledge is often built upon its use and negotiation includes an interpretive process of making meaning of exchanged ideas.

9.5 Discussion

In this chapter, we identified six practical challenges for developing a standardized assessment for CPS and outlined our general strategies for addressing them. These challenges are pragmatic in nature. It was necessary to address these challenges, so that we could develop tasks to address the important psychometric questions of reliability, validity, comparability, and fairness. We also presented our task and data collection designs along with some preliminary results from a collaborative science assessment prototype. It is worth noting that most of the challenges we mentioned so far are more from the scientific perspective. There are in addition many technical challenges when developing a CPS assessment, as it is not a trivial task to collect data from many teams in the real world. For example, the infrastructure needed for the test administration and for a flawless pairing of test takers into dyads, and the data collection and management are considerable technical challenges that were not discussed here. This study was the first we conducted to investigate the measurement of CPS skills. Despite the limitations of this exploratory study, the dataset has enabled us to map out many thought-provoking relationships. Meanwhile, we also accumulated first-hand experiences and learned many lessons during the process. The comprehensive analyses and findings from the CSAP project will be reported in our forthcoming work.

References

Adams, R., Vista, A., Scoular, C., Awwal, N., Griffin, P., & Care, E. (2015). Automatic coding procedures. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 115–132). New York, NY, USA: Springer.

Andrews, J., Kerr, D., Mislevy, R., Von Davier, A. A., Hao, J., & Liu, L. (in press). Using a simulation-based task to explore gender and cultural differences in collaboration. *Journal of Educational Measurement.*

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences, 12*(3), 307–359.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Care, E., & Griffin, P. (2014). An approach to assessment of collaborative problem solving. *Research & Practice in Technology Enhanced Learning, 9*(3), 367–388.

Cohen, E. G., Lotan, R. A., Scarloss, B. A., & Arellano, A. R. (1999). Complex instruction: Equity in cooperative learning classrooms. *Theory Into Practice, 38*(2), 80–86.

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: A meta-analysis. *Journal of Applied Psychology, 95*(1), 32–53.

Dillenbourg, P., Järvelä, S., & Fischer, F. (2009). The evolution of research on computer-supported collaborative learning. In N. Balacheff, S. Ludvigsen, T. de Jong, A. Lazonder, & S. Barnes (Eds.), *Technology-enhanced learning* (pp. 3–19). New York, NY, USA: Springer.

Dillenbourg, P., & Traum, D. (2006). Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *Journal of the Learning Sciences, 15*(1), 121–151.

Feng, S., Stewart, J., Clewley, D., & Graesser, A. C. (2015). Emotional, epistemic, and neutral feedback in autotutor trialogues to improve reading comprehension. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Proceedings of the 17th international conference on artificial intelligence in education* (pp. 570–573). New York, NY, USA: Springer.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*(3), 613–619.

Flor, M., Yoon, S.-Y., Hao, J., Liu, L., & von Davier, A. (2016). Automated classification of collaborative problem solving interactions in simulated science tasks. In P*roceedings of 11th workshop on innovative use of NLP for building educational applications* (pp. 31–41). Stroudsburg, PA, USA: Association for Computational Linguistics.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality, 37*(6), 504–528.

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills: Methods and approach*. New York, NY, USA: Springer.

Halpin, P. F., von Davier, A. A., Hao, J., & Liu, L. (in press). Measuring student engagement during collaboration. *Journal of Educational Measurement.*

Hao, J., Liu, L., von Davier, A., & Kyllonen, P. (2015). Assessing collaborative problem solving with simulation based tasks. In O. Lindwall, P. Hakkinen, T. Koschmann, P. Tchounkikine, & S. Ludvigsen (Eds.), Exploring the material conditions of learning: The computer supported collaborative learning (CSCL) conference 2015 (Vol. 1, pp. 544–547).

Hao, J., Smith, L., Mislevy, R., von Davier, A., & Bauer, M. (2016a). *Taming log files from game and simulation-based assessment: Data model and data analysis tool* (Research Report No. RR-16-10). Princeton, NJ, USA: Educational Testing Service.

Hao, J., Liu, L., von Davier, A., Kyllonen, P., & Kitchen, C., (2016b). Collaborative problem-solving skills versus collaboration outcomes: findings from statistical analysis and data mining. *Proceedings of the 9th International Conference on Educational Data Mining*.

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 37–56). New York, NY, USA: Springer.

Ho, T. K. (1995). Random decision forests. (In P*roceedings of the third international conference on document analysis and recognition* (Vol. 1, pp. 278–282). Los Alamitos, CA: IEEE Computer Society Press.

Järvelä, S., Volet, S., & Järvenoja, H. (2010). Research on motivation in collaborative learning: Moving beyond the cognitive-situative divide and combining individual and social processes. *Educational Psychologist, 45*(1), 15–27.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65*(4), 681–706.

Kerr, D., Andrews, J., & Mislevy, R. (in press). The in-task assessment framework: Extracting evidence of proficiency from in-task behavior. In A. A. Rupp & J. Leighton (Eds.), *Handbook of cognition and assessment: Frameworks, methods, and applications*. New York, NY, USA: Wiley.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *CHI '08: Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453–456). New York, NY, USA: ACM.

Koschmann, T. D. (1996). *CSCL: Theory and practice of an emerging paradigm*. New York, NY, USA: Routledge.

Kreijns, K., Kirschner, P. A., & Jochems, W. (2002). The sociability of computer-supported collaborative learning environments. *Educational Technology & Society, 5*(1), 8–22.

Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, D. (2015). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 344–359). Hershey, PA, USA: IGI Global.

Lotan, R. A. (2003). Group-worthy tasks. *Educational Leadership, 60*(6), 72–75.

Mislevy, R. J., & Riconscente, M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). New York, NY, USA: Routledge.

National Assessment of Educational Progress. (2013). *Questionnaires for students, teachers, and schools.* Retrieved July 20, 2016, from https://nces.ed.gov/nationsreportcard/bgquest.aspx

O'Neil, H. F., Jr. (Ed.). (2014). *Workforce readiness: Competencies and assessment*. New York, NY, USA: Psychology Press.

Organization for Economic Co-operation and Development. (2013). *PISA 2015 draft collaborative problem solving assessment framework*. Paris, France: Author.

Roschelle, J. (1992). Learning by collaborating: Convergent conceptual change. *Journal of the Learning Sciences, 2*(3), 235–276.

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O'Malley (Ed.), *Computer supported collaborative learning* (pp. 69–97). New York, NY, USA: Springer.

Rundgren, C.-J., Rundgren, S.-N. C., Tseng, Y.-H., Lin, P.-L., & Chang, C.-Y. (2012). Are you slim? Developing an instrument for civic scientific literacy measurement (SLiM) based on media coverage. *Public Understanding of Science, 21*(6), 759–773.

Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge-building communities. *The Journal of the Learning Sciences, 3*(3), 265–283.

Sottilare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). *The generalized intelligent framework for tutoring (gift).* Orlando, FL: US Army Research Laboratory-Human Research & Engineering Directorate (ARL-HRED).

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge (acting with technology)*. Cambridge, MA, USA: MIT Press.

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 409–426). Cambridge, UK: Cambridge University Press.

Sycara, K., Gelfand, M., & Abbe, A. (2013). *Models for intercultural collaboration and negotiation*. New York, NY, USA: Springer.

Van den Bossche, P., Gijselaers, W. H., Segers, M., & Kirschner, P. A. (2006). Social and cognitive factors driving teamwork in collaborative learning environments team learning beliefs and behaviors. *Small Group Research, 37*(5), 490–521.

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations* (Research Report No. RR-13-41). Princeton, NJ, USA: Educational Testing Service.

Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity issues in collaborative group assessment: Group composition and performance. *American Educational Research Journal, 35*(4), 607–651.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686–688.

Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., & Katz, I. R. (2014). Assessing science inquiry skills using trialogues. In S. Trausan-Matu, K. E Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent tutoring systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5–9, 2014* (pp. 625–626). New York, NY, USA: Springer.

Chapter 10
Exploring Dual Eye Tracking as a Tool to Assess Collaboration

Jennifer K. Olsen, Vincent Aleven, and Nikol Rummel

Abstract In working towards unraveling the mechanisms of productive collaborative learning, dual eye tracking is a potentially helpful methodology. Dual eye tracking is a method where eye-tracking data from people working on a task are analyzed jointly, for example to extract measures of joint visual attention. We explore how eye gaze relates to effective collaborative learning and how analysis of dual eye-tracking data might enhance analysis of other data streams. In this chapter, we identify three broad areas of analysis where dual eye tracking may enhance understanding of collaborative learning processes: (a) how eye gaze is associated with other communication measures, (b) how eye gaze is associated with features of the task environment, and (c) how eye gaze relates to learning outcomes. We present analyses in each of the three areas through joint visual attention, using a dataset of 28 fourth- and fifth-grade student dyads working on an intelligent tutoring system for fractions. By combining eye tracking, dialogue transcripts, tutor logs, and pre/post data, we show the potential of using dual eye tracking to better understand the collaborative learning process.

Keywords Collaborative learning · Intelligent tutoring system · Dual eye tracking

J.K. Olsen (✉) · V. Aleven · N. Rummel
Human Computer Interaction Institute, Carnegie Mellon University,
Pittsburgh, PA, USA
e-mail: jkolsen@cs.cmu.edu

V. Aleven
e-mail: aleven@cs.cmu.edu

N. Rummel
e-mail: nikol.rummel@rub.de

N. Rummel
Institute of Educational Research, Ruhr-Universität Bochum, Bochum, Germany

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_10

10.1 Introduction

Collaboration can be an effective way of learning; however, it is challenging to identify mechanisms of productive collaboration and to ascertain how students' actions lead to learning when working in a group. The communication between partners is likely to play a large role in the success of the group (Chi & Wylie, 2014), and there are many different processes that happen during a collaborative session that can affect learning, such as speech, joint visual attention, and tutor feedback. By analyzing these different processes separately and together, we may be able to develop a better understanding of the collaborative learning process. In this chapter, we focus on dual eye tracking, a method where eye-tracking data from collaborating partners are gathered and are analyzed jointly, to investigate whether eye movement data reveal information about collaboration that may not be readily apparent in other data streams (Jermann, Mullins, Nüssli, & Dillenbourg, 2011; Richardson & Dale, 2005). We focus on learning with an intelligent tutoring system (ITS) for fourth- and fifth-grade fractions learning that supports learning collaboratively, a feature which is atypical of ITSs (Olsen, Belenky, Aleven, & Rummel, 2014). We explore how dual eye-tracking data could be used with other data streams to analyze students' collaborative interactions. By using multiple data streams that include eye gaze, we may be able to gain insights into collaboration that would not otherwise be possible.

Research has shown that eye gaze is tied to communication, making eye tracking a promising method to use for the analysis of collaborative learning (Meyer, Sleiderink, & Levelt, 1998). Previous research has demonstrated a link between eye gaze and speech (Griffin & Bock, 2000; Meyer et al., 1998). When people hear a reference through speech, their eye gaze is likely to follow the referenced object (Meyer et al., 1998). Similarly, when people describe a picture, their eye gaze is likely to fixate on a relevant part of the picture before they describe it (Griffin & Bock, 2000). These studies show a link between speech and eye gaze that goes in both directions: eye gaze can precede the mention of an object or follow it. This same pattern occurs when people work on a task together. There is a coupling of the collaborators' eye gaze around a reference (Richardson, Dale, & Kirkham, 2007), meaning that the collaborators' gaze may fixate, at approximately the same point in time, at the object referenced in the dialogue, for example just before mentioning it and just after hearing about it. The eye gaze has a closer coupling when each of the collaborators has the same initial information and when collaborators can visually share important objects that they are referencing in speech (Jermann & Nüssli, 2012; Richardson et al., 2007), suggesting that task features influence eye gaze. The coupling of eye gaze between collaborating partners may be an indicator of interaction quality and comprehension (Jermann et al., 2011; Richardson & Dale, 2005). It also may be associated with better learning, assuming there is more comprehension and understanding from interactions with a closer coupling of eye gaze. In addition to using eye tracking as an analysis tool, eye tracking has also been used within the learning environment to signal to collaborating partners what each is looking at (Schneider & Pea, 2013). Much of the previous work using eye tracking as an analysis tool has focused on the correlation of eye gaze with speech. It is still undetermined how dual eye tracking can be used to assess the effectiveness of collaboration for learning and how it may be associated with other process data, especially within an ITS.

In this chapter, we explore three types of broad questions that can be answered by using dual eye tracking: (a) How is eye gaze associated with other communication measures? (b) How is eye gaze associated with task features? (c) How is eye gaze associated with learning outcomes? In our work, we have looked at how these three questions can be answered when students are working with an ITS. Learning with an ITS when working individually has been shown to be very successful, especially for mathematics (Ritter, Anderson, Koedinger, & Corbett, 2007; Rau, Aleven, & Rummel, 2012). In our ITS, we support collaboration through an embedded collaboration script, and we are able to study collaboration through the collection of log data, transcript data, dual eye-tracking data, and pretest/posttest data. By answering the questions listed above, we may develop a better understanding of how the different features of the learning process relate and have an impact on learning for students who are collaborating.

Multiple measures can be gathered through dual eye tracking to understand eye gaze during collaboration. In this chapter, we focus on one such measure, joint visual attention, which measures the coupling of eye gaze as the relative amount of time two collaborating students look at the same area at the same time. Using a dataset gathered from fourth- and fifth-grade students working on an ITS for fractions learning, we explored a specific question in each of the three broad areas outlined above. These exploratory analyses demonstrate the potential of combining dual eye tracking and other data streams to analyze collaborative learning.

10.2 Methods

10.2.1 Experimental Design and Procedure

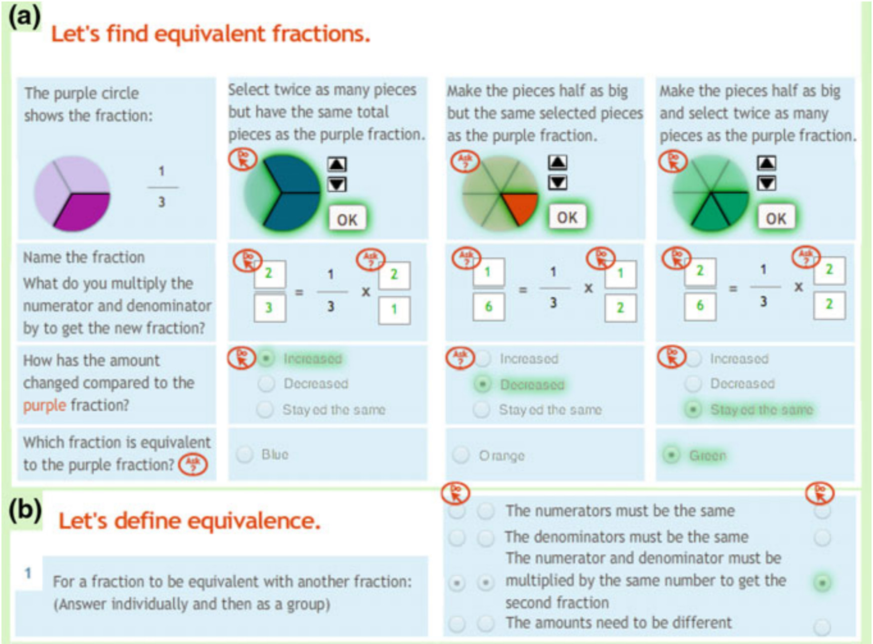
Our dataset involves 14 fourth-grade and 14 fifth-grade student dyads from a larger study in which we tested a hypothesis about differential benefits of collaborative versus individual learning (Olsen et al., 2014). The current chapter focuses, not on that hypothesis, but on the use of dual eye tracking in collaborative learning research. The dyads were engaged in a problem-solving activity using a networked collaborative ITS while communicating through audio only, using Skype. The study used a between-subjects design. Each teacher paired the students participating in the study, matching students who would work well together and who had similar, but not equivalent, math abilities. The pairs were then randomly assigned to either work collaboratively or individually and on either a procedurally-oriented or a conceptually-oriented problem set. (In this chapter, we present data from the dyads only.) Each dyad worked with the tutor for 45 min in a lab setting at their school.

The morning before working with the tutor and the morning after working with the tutor, students were given 25 min. To complete a pretest or posttest individually on the computer to assess their learning. Although the lab was set up in the school, we were able to collect dual eye-tracking data, dialogue data, and tutor log data in addition to the pretest and posttest measures.

10.2.2 Tutor Design

The dyads in our study used a tutoring system for fractions learning that we developed using the Cognitive Tutor Authoring Tools (Aleven, McLaren, Sewall, & Koedinger, 2009; Aleven et al., 2016; Olsen et al., 2014). This tutor consisted of two problem sets, one targeting procedural knowledge and one targeting conceptual knowledge about fraction equivalence. Procedural knowledge is the knowledge about the steps needed to solve a problem and the ability to execute these steps in the correct order (Rittle-Johnson, Siegler, & Alibali, 2001). Conceptual knowledge is the knowledge of how the different elements of the domain are interrelated (Rittle-Johnson et al., 2001). Each dyad in the study worked on only one of these two problem sets, given the study goals mentioned above. Within each problem, the tutor provided standard ITS support, such as prompts for steps, next-step hints, and step-level feedback that allowed the problem to adapt to the student's problem-solving strategy (VanLehn, 2011). For the collaboration, the ITS support mentioned above was combined with embedded collaboration scripts. Each subgoal, a group of related steps (e.g., finding the factors of 9) in the problem was revealed one at a time. Each student had his or her own view of the collaborative tutor on separate computers that allowed the students to have a shared problem space and synchronously work while being able to see slightly different information and to take different actions.

In addition to providing step-level guidance to students within each problem, the tutor was designed to support effective collaboration between students in three different ways. First, for many steps, the students were assigned roles (see Fig. 10.1); this process has been shown to be an effective collaboration scripting feature (King, 1999). Roles support collaboration by assigning students certain tasks within the given problem. This provides the students with guidance for their own responsibilities and with an understanding of their partner's responsibilities. In our tutors, on steps with roles, one student was responsible for entering the answer and the other was responsible for asking questions of his or her partner and providing help with the answer. The tutor indicated the current role for the students through the use of icons on the screen (see Fig. 10.1). A second way that collaboration was supported was by providing students with information they were responsible for sharing with their partner, *individual information* (Slavin, 1996). The students were each provided with a different piece of information needed for the solution to the problem; thus they needed to share this information with their partner as indicated by a "Share" icon (see Fig. 10.2). The final feature that was



(a) Let's find equivalent fractions.

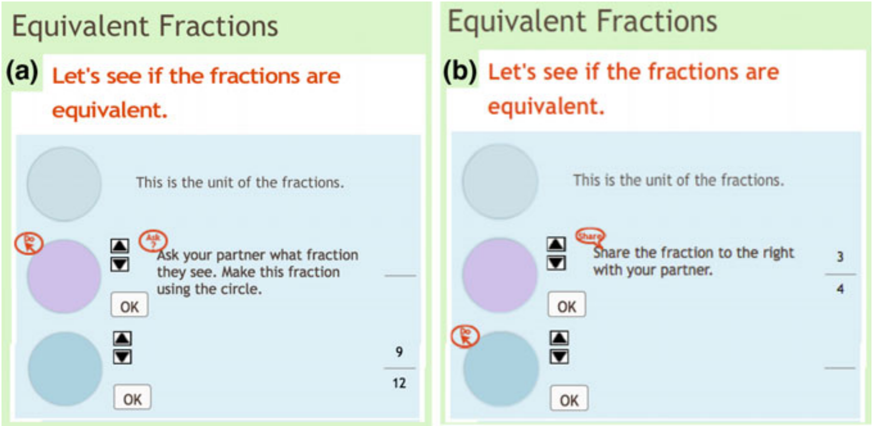
The purple circle shows the fraction: $\frac{1}{3}$	Select twice as many pieces but have the same total pieces as the purple fraction. OK	Make the pieces half as big but the same selected pieces as the purple fraction. OK	Make the pieces half as big and select twice as many pieces as the purple fraction. OK
Name the fraction What do you multiply the numerator and denominator by to get the new fraction?	$\frac{2}{3} = \frac{1}{3} \times \frac{2}{1}$	$\frac{1}{6} = \frac{1}{3} \times \frac{1}{2}$	$\frac{2}{6} = \frac{1}{3} \times \frac{2}{2}$
How has the amount changed compared to the purple fraction?	● Increased ○ Decreased ○ Stayed the same	○ Increased ● Decreased ○ Stayed the same	○ Increased ○ Decreased ● Stayed the same
Which fraction is equivalent to the purple fraction?	○ Blue	○ Orange	● Green

(b) Let's define equivalence.

1 For a fraction to be equivalent with another fraction: (Answer individually and then as a group)

○ ○ The numerators must be the same ○
○ ○ The denominators must be the same ○
● ● The numerator and denominator must be multiplied by the same number to get the second fraction ●
○ ○ The amounts need to be different ○

Fig. 10.1 For the display of a single student, section A shows an example of roles for each subgoal of the problem where the "Do" icon indicates the student is responsible for entering the answer while the "Ask" icon indicates the student is responsible for asking questions and helping to find the correct answer. Section B shows an example of cognitive group awareness



Equivalent Fractions

(a) Let's see if the fractions are equivalent.

This is the unit of the fractions.

Ask your partner what fraction they see. Make this fraction using the circle.

OK

OK $\frac{9}{12}$

Equivalent Fractions

(b) Let's see if the fractions are equivalent.

This is the unit of the fractions.

Share the fraction to the right with your partner. $\frac{3}{4}$

OK

OK

Fig. 10.2 An example of individual information where Student A is responsible for making the fraction that the partner, Student B, (*right*) has in symbolic form and must share

used to support collaboration was *cognitive group awareness,* such that knowledge that each student had in the group was made known to the group (Janssen & Bodemer, 2013). This feature was implemented on steps where the students needed to extract a pattern from the earlier steps in the problem. Each student was given an opportunity to answer a question individually before the students were shown each other's answers and asked to provide a consensus answer (see Fig. 10.1).

10.2.3 Data and Dependent Measures

A computer-based test was developed to closely match the target knowledge covered in the tutors. The test comprised five procedural and six conceptual test items, based on pilot studies with similar materials. Two isomorphic sets of questions were developed, and there were no differences in performance on the test forms, $t(79) = 0.96$, $p = 0.34$. The presentation of these forms as pretests and posttests was counterbalanced. Between the pretest ($M = 2.06$, $SD = 1.25$) and the posttest ($M = 2.56$, $SD = 1.05$) for conceptual knowledge, there were significant learning gains, $F(1, 25) = 7.66$, $p = 0.010$, but there were no learning gains on procedural knowledge from pretest ($M = 0.70$, $SD = 0.77$) to posttest ($M = 0.87$, $SD = 0.84$), $F(1, 25) = 1.13$, $p = 0.296$ (Belenky, Ringenberg, Olsen, Aleven, & Rummel, 2014).

In addition, to pretest and posttest measures, we also collected process data, including dual eye-tracking data, tutor log data, and dialogue data. We collected eye-tracking data using two SMI Red 250 Hz infrared eye-tracking cameras. We recorded each student's eye movement separately, synchronized the eye-tracking logs of the students in each dyad, and analyzed the fixation data of the students in a dyad jointly. We calculated a measure of joint visual attention through gaze recurrence (Belenky et al., 2014; Marwan, Romano, Thiel, & Kurths, 2007). Gaze recurrence is the proportion of time that collaborating students fixate their gaze simultaneously at the same location. In other words, it is the proportion of time that the students' eye gazes are coupled. To calculate the joint visual attention from the gaze data, we used gaze recurrence with a distance threshold of 100 pixels to approximate the percentage of time that students were looking at the same thing at the same time. This distance threshold was chosen to align with prior research (Jermann et al., 2011) and is close to the size of many of the interface elements.

The log data captured the time-stamped transactions that the students took with the ITS. These include attempts at solving each step and their request for hints; the log data also includes the tutor's responses, including whether attempts at solving were correct, what knowledge components they involved, and what errors were made, as well as any hint and feedback messages that the tutor presented to the students.

We transcribed the students' dialogues and coded the transcript data using a rating scheme with four categories: *interactive dialogue*, *constructive dialogue*,

Table 10.1 Rating scheme category definitions and mapping to the ICAP framework

Type of talk	Overt actions	ICAP framework
Interactive dialogue	Discussing an answer, coconstruction, soliciting help or confirmation of agreement	Joint dialogue
Constructive dialogue	Guessing as a group, argumentation without explanation, agreeing with partner without adding on	Individual dialogue
Constructive monologue	Self-explanation	Individual dialogue
Other	Telling the answer, work coordination, active reading, and off-topic talk	

constructive monologue, and *other*. We developed this rating scheme to align with the ICAP (Interactive Constructive Active Passive) framework (Chi, 2009) and to distinguish between the different types of talk, ranked on how conducive we hypothesize they would be for learning. For our analysis, we focused on the interactive dialogue, which aligns with ICAP's joint dialogue pattern (Chi, 2009) and is hypothesized to be more conducive to learning than other types of talk and dialogue. In interactive dialogue, students engage in actions such as sequential construction and co-construction. In sequential construction, each student allows his or her partner to finish the turn before adding additional information, while in co-construction, students do not wait for their partner's turn to finish but instead finish the partner's thought. Our rating scheme was developed to look at utterances associated with each subgoal (i.e., a group of related steps within a tutor problem) to account for the interactions between the students. An interrater reliability analysis was performed to determine consistency among raters (kappa = 0.72) (Table 10.1).

10.3 Research Questions and Analysis

We now illustrate how we used dual eye-tracking data, in conjunction with other data sources, to study collaborative learning processes and their relations with learning. We focus on each of our three questions in turn.

10.3.1 Relation Between Eye Gaze and Dialogue

The first broad area of analysis is how eye gaze is associated with other communication measures, specifically our coding of the dialogue data. By understanding the association between eye gaze and other measures of communication, we may begin to understand how eye gaze and dialogue interrelate, as well as where dual eye tracking might provide information about the collaboration that dialogue data does not reveal. Specifically, we investigated how joint visual attention differs between subgoals without talk and subgoals with talk. Based on previous work, we hypothesized that subgoals with talk would have a higher level of joint visual attention than subgoals with no talk. As mentioned, research has shown that talk is coupled with eye gaze; in particular, speech can guide visual attention (Meyer et al., 1998).

We extended these prior analyses by asking whether relations between eye gaze and dialogue vary depending on what happens at the problem-solving level, namely, whether students commit problem-solving errors or solve steps correctly, as indicated by tutor feedback. As mentioned, when students enter their attempts at solving a step into the tutor interface, the software responds by providing color-coded correctness feedback, with green indicating correct answers and red indicating incorrect answers. Because errors are often viewed as learning opportunities (Ohlsson, 1996), it is interesting to ask whether, in collaborative learning scenarios, they tend to be moments of particularly intense collaboration. Suggestive of that notion, in our dataset, we found that subgoals with errors have higher frequency of talk (Olsen, Rummel, & Aleven, 2015). Here we asked whether errors may show interesting relations with eye gaze and whether errors modify the relation between eye gaze and dialogue measures. Not only do errors have a clear visual manifestation on the screen, as they serve as an external record of the last step entered, made especially salient by the tutor's red feedback, but also as students discuss the error, their eye gaze may fixate on the object of discussion (i.e., the error; Richardson et al., 2007). Therefore, we hypothesized that subgoals on which an error occurred would have a higher level of joint visual attention than subgoals where no error occurred.

To address these hypotheses, we investigated how joint visual attention differs between subgoals with talk and subgoals without talk. We also explored whether or not there is an interaction between errors and talk, regarding the level of visual attention, such that the greatest level of joint visual attention is found for subgoals with talk and errors (see Fig. 10.3). We used a hierarchical linear model with two nested levels to analyze how the talk during subgoals related to our dependent variable of joint visual attention. At Level 1, we modeled whether talk occurred and whether one or more errors occurred for the subgoals as our independent variables. At Level 2, we accounted for random dyad differences. We found no effect of errors on joint visual attention, so we removed the error variable from the model as an independent variable. We found greater joint visual attention for subgoals that had talk ($M = 0.25$, $SD = 0.13$) versus those that did not ($M = 0.22$, $SD = 0.14$, $t(1705) = 2.66$, $p < 0.001$), $\omega^2 = 0.06$, showing a coupling between talk and joint visual attention that extends previous results to younger learners working in an ITS environment. However, we did not find support for our hypothesis that the presence of errors has an impact on joint visual attention.

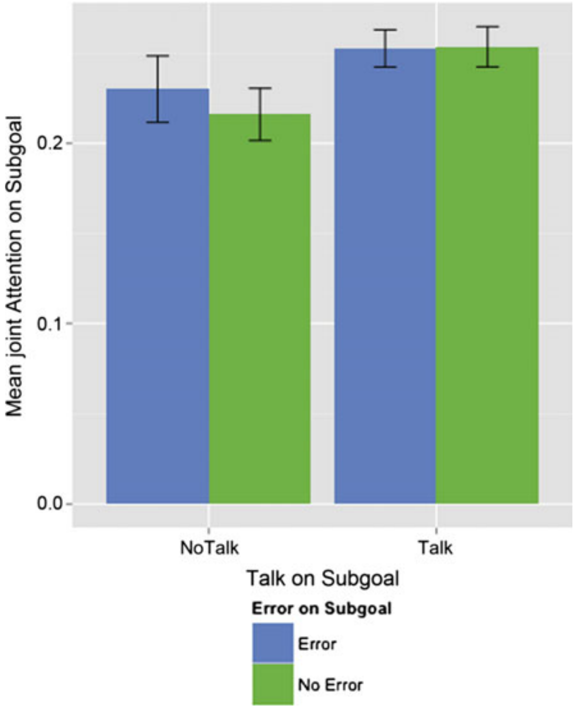


Fig. 10.3 Joint visual attention (with standard errors) as a function of whether there was talk on the given subgoal and whether there were errors on the given subgoal

10.3.2 Relation Between Eye Gaze and Tutor Support for Collaboration

The second broad area of analysis is how eye gaze is associated with features of the task environment, in our case, the design of the tutor interface to support collaboration. The tutor provided a different interface for different problem types, since the interface is designed to make the steps of the problems explicit for the students. Cutting across these different interfaces, however, are the three tutor features that support collaboration, described above (roles, individual information, and cognitive group awareness). We focused on these collaborative features and how they might affect joint eye gaze measures. By analyzing the association between eye gaze and different task features, we can begin to understand the impact that task features can have at the process level, beyond what can be abstracted from student dialogue. As well, this investigation reveals the degree to which the support for collaboration provided by the tutor manifests itself in joint visual attention. Based on previous work, we hypothesized that subgoals supported through individual information would have the lowest joint visual attention, compared to subgoals with the two other collaborative features, since there is no joint reference for the students on the screen (Jermann & Nüssli, 2012). We did not have an expectation for whether the

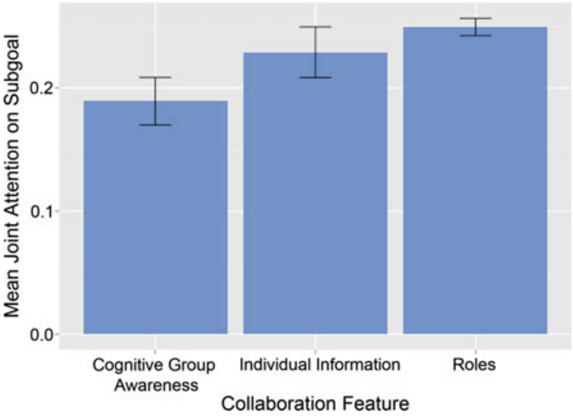


Fig. 10.4 Joint visual attention (with standard errors), as function of the collaborative features present

cognitive awareness feature and the roles feature would lead to differences in joint visual attention.

To investigate the association between collaboration features and joint visual attention, a hierarchical linear model with two nested levels was used to analyze how collaboration features relate to joint visual attention as the dependent variable. At Level 1, we modeled the type of collaboration support of the subgoals, along with the talk type to control for this covariate as the independent variables. At Level 2, we accounted for random dyad differences. We found that the joint visual attention for subgoals that were supported through cognitive group awareness (M = 0.19, SD = 0.11) was lower than that for subgoals supported through roles (M = 0.25, SD = 0.14), $t(1705)= -4.19$, $p <0 .001$, $\omega^2 = 0.10$, indicating that how the task environment of supports for collaboration seems to have an impact on joint visual attention (see Fig. 10.4). These results do not support our hypothesis that individual information would have the lowest joint visual attention.

10.3.3 Relation Between Eye Gaze and Learning Outcomes

The third broad area of analysis is how eye gaze is associated with learning gains. Within this area, we investigated how joint visual attention correlates with learning gains for conceptual and procedural knowledge, as measured by pretest and posttest. Our initial hypothesis was that joint visual attention would be associated with greater learning gains, more strongly so for students working with conceptually-based problems. This hypothesis tests the intuition that if joint eye gaze is an indicator of productive collaborative learning processes, then it should correlate with the learning outcomes of these processes. This notion finds some support in earlier research that found a positive relationship between understanding and joint eye gaze (Richardson & Dale, 2005). However, that study did not distinguish between the types of knowledge that were being acquired. Our hypothesis takes into account the types of knowledge the students are targeting and is informed by prior work that found that collaboration can be more beneficial for learning conceptual knowledge than for learning procedural knowledge (Mullins, Rummel, & Spada, 2011). Therefore, we predicted the correlation between joint eye gaze and conceptual learning gains to be stronger than that between joint eye gaze and procedural learning gains.

To investigate this question, we computed a hierarchical linear model with two nested levels to analyze how posttest scores (dependent variable) correlated to joint visual attention as the independent variable, while controlling for the pretest score as a covariate. At Level 1, we modeled the joint visual attention and the pretest scores. At Level 2, we accounted for random dyad differences. The joint visual attention was calculated for each dyad for the entire 45-minute session. We found, as hypothesized, that joint visual attention significantly predicts conceptual posttest scores when controlling for conceptual pretest score. However, contrary to our expectations, this effect was confined to the students in the procedural condition, $t(11) = 2.30$, $p = 0.04$, $\omega^2 = 0.57$ (see Fig. 10.5). Recall that these students solved problems targeting procedural knowledge of fractions only. There was no significant correlation between procedural learning and joint eye gaze (see Fig. 10.6).

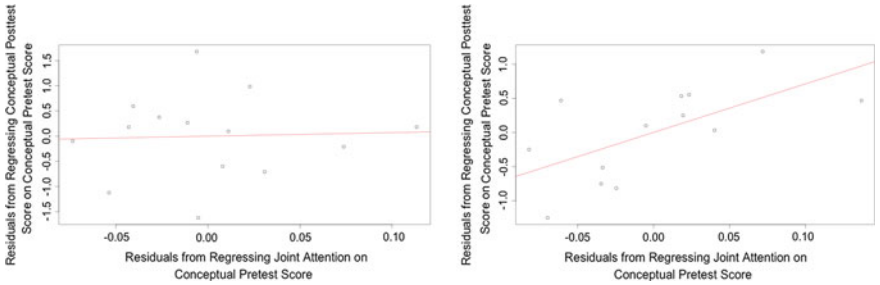


Fig. 10.5 Partial correlation of joint visual attention with posttest scores on conceptual knowledge for students in the conceptual condition (*left*) and the procedural condition (*right*)

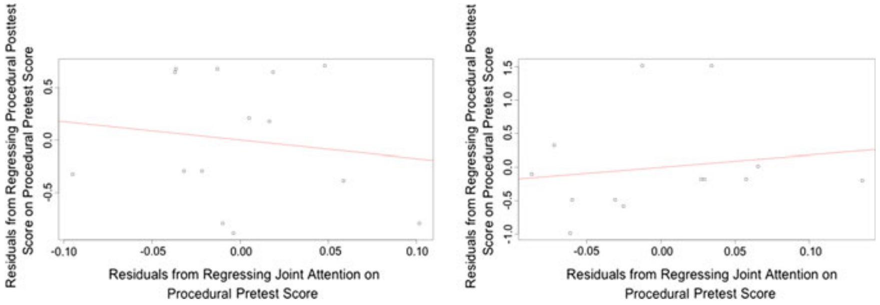


Fig. 10.6 Partial correlation of joint visual attention with posttest scores on procedural knowledge for students in the conceptual condition (*left*) and in the procedural condition (*right*)

These results thus provide partial support for our hypothesis. We note that these results are consistent with preliminary findings based on a subset of the data (Belenky et al., 2014).

10.4 Discussion

Our project studies learning in a collaborative tutoring environment. One of its aims is to utilize multiple data sources, including dual eye tracking, to understand relationships between joint visual attention, dialogue, problem-solving performance, and learning. In this chapter, we have explored different roles that dual eye tracking, in combination with other data sources, may play in understanding these relations. It may help advance our understanding of how dual eye tracking can be useful in understanding collaborative learning.

Although the correspondence of eye gaze with speech has been studied before, it is still undetermined whether and how dual eye tracking can be used to assess the effectiveness of collaboration for learning and how it is associated with other process data. Nor, to the best of our knowledge, has dual eye tracking been used before to study mathematics learning of elementary school students, supported by ITS software. In this paper, we have explored the importance of eye gaze for collaborative learning analysis by presenting three different areas of analysis using dual eye-tracking data. These areas provide a broad structure and illustrate the potential of dual eye tracking, especially when used in conjunction with other data streams. These areas of analysis have provided some interesting, if sometimes unexpected, findings that warrant further investigation.

To what degree does dual eye tracking contribute to understanding collaborative learning processes? Through our analysis, we found that subgoals where talk occurs have a higher level of joint visual attention than subgoals without talk, extending previous work (Richardson et al., 2007) to younger learners and to working in an ITS environment. This result suggests, in line with prior work, that speech can help coordinate joint visual attention, for example by referencing items on the screen. Interestingly, it can do so even in a task environment that may already drive eye gaze to certain areas of the screen—in the tutor, there is step-by-step guidance and subgoals are revealed one at a time, which may provide a strong suggestion to the collaborating partners of where to place their attention. Contrary to our hypothesis, we did not find greater joint visual attention on subgoals where students made errors. Apparently, if errors are occasion for more frequent or more intense collaboration, as our analysis of the speech data suggests (Olsen et al., 2015), this effect does not manifest itself through increased joint visual attention. It may be that neither tutor feedback marking the error in red on the screen, nor discussion of errors with a partner, caused greater joint visual attention than answering the item originally. It is possible that errors may not lead to greater collaboration, although that would be inconsistent with the speech data. Alternatively joint visual attention, especially when considered at the subgoal level, may be too temporally coarse-grained as a measure of collaboration. Analyzing the joint visual attention immediately after an error (i.e., at a finer temporal grain size) may provide a better indication of the effect of errors on joint visual attention.

In addition, we found differences in the level of joint visual attention associated with three tutor features designed to support collaboration, albeit in somewhat unexpected ways. Contrary to our expectation, subgoals supported through cognitive group awareness had a lower level of joint visual attention compared to those supported through roles. We must note that this conclusion is tentative, as the analysis does not fully separate the effect of the specific type of fraction subgoal (e.g., whether students are trying to understand factors versus the notion that the numerator and denominator are multiplied by the same number for equivalent fractions) from that of the specific type of collaboration support. Not all problem types were crossed with all support types.

Nonetheless, it is interesting to ask why subgoals with the cognitive awareness feature may have lower joint visual attention than those with roles. Recall that on subgoals supported through cognitive group awareness, students first answer a multiple choice question individually (see Fig. 10.1); they then get to see their partner's answer, and then (presumably after discussing their individual answers, at least if they differ) provide a consensus answer. Although students may have their attention on this same question, they may not be looking in the same area of the question because they are trying to understand the visual information that came from their partner. It may be, as well, that the students do not discuss the group answer before entering it, so that a common verbal reference that would guide the eye gaze is lacking. More temporally fine-grained analysis of joint eye gaze may help shed light on this somewhat speculative interpretation.

Alternatively, it may be worthwhile to consider whether there is greater joint visual attention (perhaps coupled with more talk) when the partners' individual answers diverge. On the other hand, when the students were supported through roles, they may have been able to follow along as their partner submitted an answer to a step, which would lead to a higher level of joint visual attention. When there is little talk, visual attention is the key way that the partner would know when the solution to the step has been entered to the problem, by watching the screen. Here again, more fine-grained analyses of eye-tracking data may help.

Finally, it is important to look at correlations between joint eye gaze and learning outcomes, as these correlations would provide support for joint visual attention as an indicator of the degree to which the students might be collaborating productively (Jermann et al., 2011; Richardson & Dale, 2005). As hypothesized, we found joint visual attention to be a significant predictor of conceptual posttest scores. Contrary to our hypothesis, this correlation was found only in the procedural condition, in which students solved problems aimed at supporting procedural learning. Also, contrary to our hypothesis, we found no correlation between procedural learning and joint eye gaze. Combined with our finding of learning gains for conceptual knowledge, we might infer that collaboration and joint visual attention may be important for conceptual knowledge, specifically when it is not being directly supported. When conceptual knowledge is already supported in the tutor, there may be no additional gain for students to be working together, and they may have less joint eye gaze. The difference in correlation might also be due to the way the problems are developed. The conceptual problems tend to be much more text heavy than the procedural problems, which may lead to having less overall joint eye gaze for conceptual problems.

Our results show the potential of using dual eye tracking to better understand collaboration, especially when it is used in conjunction with other data streams, although there are some limitations with our small sample size and it is unclear how our results would generalize outside of our dataset. However, our analyses do suggest that dual eye tracking can reveal additional information not evident in other data streams, and that analysis with other data streams can help guide the process of considering tentative interpretations based on eye-tracking data. We see this in all three of our different analyses. For our analysis of the correlation between talk on a subgoal and joint eye gaze, we showed that there was a relationship within our dataset, which provides information beyond just student speech, that students might be connecting problem features that they see through their dialogue. For the collaborative features used in the tutors, we gained further insights on which features may have an impact on collaboration by analyzing how joint eye gaze differed between the different features. In terms of learning, we also found a correlation of conceptual learning with joint eye gaze for students working procedurally. This may provide some insights about when collaboration may be appropriate for students during the learning process. Together, all of these analyses show the benefit of analyzing educational data in conjunction with joint eye gaze.

For future work, we would like to expand the three areas of analysis around dual eye tracking beyond joint visual attention. There are other measures, such as areas of interest (AOIs) analyses and gaze patterns that would be of interest in each of the three areas and can be measured through dual eye tracking. These different measures of eye gaze would not only provide additional ways of comparing collaboration within groups by looking at AOIs and gaze patterns that occur for partners at the same time, but would also allow the comparison to students working individually to see how collaboration affects the learning process. For example, an AOI analysis might help us distinguish between joint visual attention on areas with text versus graphical representations, and whether patterns of distributing attention between these representations differ between students working individually and those working collaboratively. In addition, in our analyses so far, we have analyzed joint visual attention at the subgoal level and the dyad level, but analysis at additional grain sizes, such as a few seconds around errors and the problem level, would allow us to address a wider range of questions.

In this chapter, we have shown that dual eye tracking can be combine with other process measures to shed light on the mechanisms of the collaborative learning process that may otherwise not be accessible. By understanding how these different data streams relate to one another, we can use a mix of the data streams in future work to better understand the dyad's learning outcomes and how the individual contributes to that learning. In this chapter, we looked at joint eye gaze with overall learning gains, but there are other process measures as well that may be used along with joint eye gaze to provide a more complete picture. The group answers of the team are recorded within the tutor logs, providing a measure of the dyad's process within the tutor. By combining the tutor logs with other process data, such as speech and eye gaze, we may be able to measure individual contributions to the group by understanding which student may have suggested an answer and whether a certain student is leading the discussion by leading the eye gaze. By combing the different data streams, we may be able to better understand the interplay between students that leads to a successful collaboration and beneficial learning.

Acknowledgments We thank the CTAT team, Michael Ringenberg, Daniel Belenky, and Amos Glenn, for their help. This work was supported by Graduate Training Grant #R305B090023 and by Award #R305A120734, both from the U.S. Department of Education (IES).

References

Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education, 19*(2), 105–154.

Aleven, V., McLaren, B. M., Sewall, J., van Velsen, M., Popescu, O., Demi, S., et al. (2016). Example-tracing tutors: Intelligent tutor development for non-programmers. *International Journal of Artificial Intelligence in Education, 26*(1), 224–269.

Belenky, D. M., Ringenberg, M., Olsen, J., Aleven, V., & Rummel, N. (2014). Using dual eye-tracking to evaluate students' collaboration with an Intelligent Tutoring System for elementary-level fractions. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 176–181). Austin, TX: Cognitive Science Society.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1,* 73–105.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11* (4), 274–279.

Janssen, J., & Bodemer, D. (2013). Coordinated computer-supported collaborative learning: Awareness and awareness tools. *Educational Psychologist, 48*(1), 40–55.

Jermann, P., Mullins, D., Nüssli, M. A., & Dillenbourg, P. (2011). Collaborative gaze footprints: Correlates of interaction quality. In H. Spada, G. Stahl, N. Miyake, & N. Law (Eds.), *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL2011 Conference Proceedings* (Vol. 1, No. EPFL-CONF-170043 pp. 184–191). Hong Kong, China: International Society of the Learning Sciences.

Jermann, P., & Nüssli, M. A. (2012). Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. In Association for Computing Machinery (Ed.), *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 1125–1134). New York: Association for Computing Machinery.

King, A. (1999). Discourse patterns for mediating peer learning. In A.M. O'Donnell & A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 87–117). Mahwah, NJ: Lawrence Erlbaum Associates.

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports, 438,* 237–329. doi:10.1016/j.physrep.2006.11.001.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition, 66,* B25–B33.

Mullins, D., Rummel, N., & Spada, H. (2011). Are two heads always better than one? Differential effects of collaboration on students' computer-supported learning in mathematics. *International Journal of Computer-Supported Collaborative Learning, 6*(3), 421–443.

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review, 103*(2), 241–262.

Olsen, J., Belenky, D., Aleven, V., & Rummel, N. (2014). Using an intelligent tutoring system to support collaborative as well as individual learning. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems, ITS 2014* (pp. 134–143). Berlin: Springer.

Olsen, J. K., Rummel, N., & Aleven, V. (2015). Finding productive talk around errors in intelligent tutoring systems. In O. Lindwall, P. Häkkinen, T. Koschmann, P. Tchounikine, & S. Ludvigsen (Eds.), *Exploring the Material Conditions of Learning: Proceedings of the International Conference on Computer Supported Collaborative Learning 2015* (Vol. 2, pp. 821–822). Gothenberg, Switzerland: International Society of the Learning Sciences.

Rau, M. A., Aleven, V., & Rummel, N. (2012). Sense making alone doesn't do it: Fluency matters too! ITS support for robust learning with multiple representations. In S. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 174–184). Berlin/Heidelberg: Springer.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29,* 1045–1060.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science, 18*(5), 407–413.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*(2), 249–255.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346–362.

Schneider, B., & Pea, R. (2013). Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning, 8*(4), 375–397.

Slavin, R. E. (1996). Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology, 21*(1), 43–69.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.

Chapter 11
Multimodal Behavioral Analytics in Intelligent Learning and Assessment Systems

Saad M. Khan

Abstract As the boundary blurs between the real and the virtual in today's learning environments, there is a growing need for new assessment tools that capture behavioral aspects key to evaluating skills such as problem solving, communication, and collaboration. A key challenge is to capture and understand student behavior at fidelity sufficient to estimate cognitive and affective states as they manifest through multiple media, including speech, body pose, gestures and gaze. However, analyzing each of these modalities in isolation may result in incongruities. In addition, the affective states of a person show significant variations in time. To address these technical challenges, this paper presents a framework for developing hierarchical computational models that provide a systematic approach for extracting meaningful evidence from noisy, unstructured data. This approach utilizes multimodal data, including audio, video, and activity log files and models the temporal dynamics of student behavior patterns. To demonstrate the efficacy of our methodology, we present two pilot studies from the domains of collaborative learning and in vivo assessments of nonverbal behavior where this approach has been successfully implemented.

Keywords Machine learning · Multimodal fusion · Hierarchical processing models

11.1 Introduction

To be successful in today's rapidly evolving, technology-mediated world, students must not only possess strong skills in areas such as reading, math, and science, but they must also be adept at 21st-century skills such as critical thinking, communication, problem solving, persistence, and collaboration (Farrington et al., 2012). These skills have been demonstrated to improve learning outcomes and are being

S.M. Khan (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: skhan002@ets.org

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_11

rapidly incorporated in a number of high-stakes standardized assessment systems (Smarter Balanced Assessment Consortium, n.d.). However, the assessment of skills such as collaboration and communication is difficult because often it involves understanding the process used to arrive at a conclusion rather than simply the end product (Bejar, 1984; Romero & Ventura, 2007). Analyzing these processes requires tracking not only the cognitive processes but also noncognitive behaviors, for example, motivation, self-control, and emotional and affective states that influence interpersonal interactions. In addition, much of the infrastructure of assessment design has come of age around traditional multiple-choice tests and self-reports. In contrast, educational simulations and games provide opportunities to expose students to authentic educational tasks and allow them to interact with and explore complex representations of serious academic content (Fisch, 2005; National Research Council, 2011). They do so in a manner that is amenable to capturing rich process data in vivo, that is, during the execution of a task involving collaboration, problem solving, and other complex tasks. These data can be multimodal, that is, they can include multiple sensory modalities such as audio, video, and 3D (using depth-sensing devices such as Microsoft Kinect), in addition to traditional forms of computer interaction data such as mouse click streams and keystrokes. The key advantage of using such multimodal data is that it enables high fidelity sensing and tracking of a user's cognitive and noncognitive states, which would otherwise be missed in traditional log files. However, extracting relevant features from these data that can be used as evidence to infer competency in complex constructs such as collaboration is a significant technical challenge for a number of reasons. First, the raw time series of multimodal data often does not have any direct semantic meaning and may not be interpretable by humans as such. As mentioned earlier, it may constitute simulation log files, audio, and visual data, which, without sophisticated computational models, cannot be analyzed for meaningful information. Second, building pattern recognition approaches to detect and recognize sequences and combinations in raw data requires "training data" that may not be readily available. And finally, the inferences and corresponding interpretations from raw multimodal data may contain information at vastly different levels of semantic meaning and abstraction that may not be easily combined in a scoring model, for example, specific facial expressions versus turn taking or user's level of engagement.

This paper is intended to provide a framework and methodology to design and develop computational models that enable analysis of noisy, unstructured, multimodal data for the assessment of complex constructs such as collaboration and communication. Specifically, this paper describes a hierarchical data processing and inference methodology that can help bridge the gap between the raw, low-level multimodal data and the measurement of high-level constructs. To illustrate the efficacy of such a methodology, two example pilot studies are presented where such an approach was implemented to study collaborative learning and in vivo measurement of nonverbal behavior using wearable sensors.

11.2 Multimodal Analytics for Studying Student Behavior

Simulations and games in computerized educational environments offer an exciting new paradigm to assess knowledge, skills, and abilities that are difficult to capture with traditional measurement tools such as structured tests and multiple choice items. Such computerized educational environments enable powerful audiovisual interfaces that can be utilized to analyze student's actions, behaviors, and indeed their process in solving the problem, rather than just their final products. Of particular interest are moment-by-moment student affective and cognitive states and how these are related with task performance and learning outcomes in general (D'Mello & Graesser, 2012; Whitehill, Serpell, Lin, Foster, & Movellan, 2014).

A key advantage of using computerized educational environments is that they can enable gathering of rich multimodal data in the form of video streams, audio streams, and simulation log files. These data can be processed and analyzed using multimodal analytics to study performance at individual and group levels. The term *multimodal analytics* (Amer, Siddiquie, Khan, Divakaran, & Sawhney, 2014; Morency, de Kok, & Gratch, 2010; Siddiquie, Khan, Divakaran, & Swahney, 2013) refers to the use of advanced sensor technologies and machine learning systems to track and understand human behaviors. It promises a paradigm shift in learning and assessment that can afford rich, automated, and grounded inferences about human performance from large amounts of multiple sensory data, for example, audio and video. However, developing computational models that can extract meaningful features indicative of performance and skills from the raw, low-level multimodal data is a significant technical challenge. In contrast, when human observers rate task performance, they are quite naturally integrating information from both what they have seen (visual) and what they have heard (auditory). Moreover, the observers' brains translate the visual data into information about body postures, facial expressions, and actions taken. The auditory data are translated into meaningful communication, multiperson verbal exchanges, and tone-of-voice cues. These features are further combined to allow the observers to make judgments about the emotional states, social skills, and technical competencies of the individual performers.

11.2.1 Hierarchical Inference Framework

To address the challenges outlined above, our approach is to build a hierarchical processing and inference framework. As illustrated in Fig. 11.1, raw multimodal data form the first layer of the framework. Data are captured using a multitude of sensors, including audio, video, 3D, and even simulation log files. These data are preprocessed to extract machine features, for example, histogram of oriented gradients (HOG) from visual data, Euler angles from 3D skeleton data, and Mel-frequency cepstral coefficient (MFCC) features from audio data, among others.

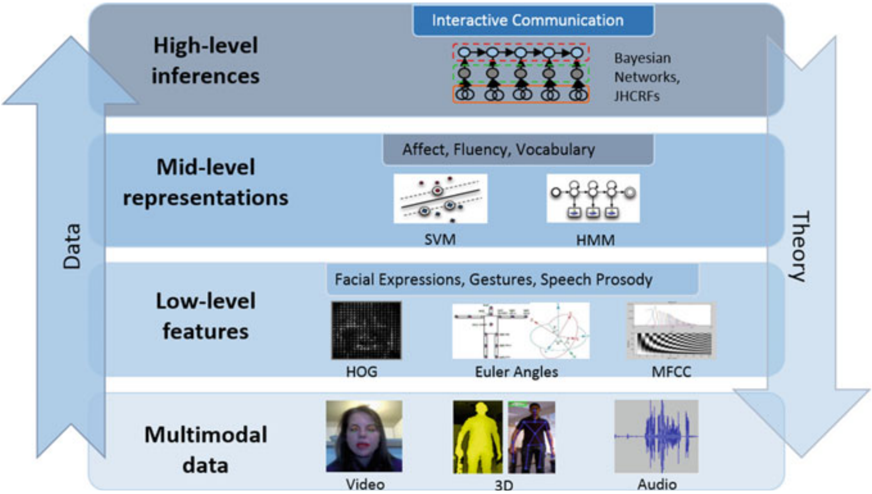

High-level inferences
Interactive Communication
Bayesian Networks, JHCRFs
Mid-level representations
Affect, Fluency, Vocabulary
SVM
HMM
Low-level features
Facial Expressions, Gestures, Speech Prosody
HOG
Euler Angles
MFCC
Multimodal data
Video
3D
Audio
Data
Theory

Fig. 11.1 Our framework to bridge the gap between low-level digital data and the measurement of complex constructs. *HMM* hidden Markov model, *HOG* histogram of oriented gradients, *JHCRF* joint hidden conditional random fields, *MFCC* mel-frequency cepstral coefficient, *SVM* support vector machine

We call such machine features *low-level features*, and they reside in the second layer of the hierarchical framework. The output of this layer is descriptive features that may have semantic meanings, such as facial expressions, gestures, or speech prosody. Such descriptive features termed *mid-level representations* reside the next level up in the hierarchical framework.

In this layer, the temporal dynamics of low-level features and mid-level representations are modeled to generate holistic measures of human behavioral states such as affect, engagement, and flow. At the top level of the hierarchy reside the features that make up a theoretical model representing the construct of interest such as communication competency or collaborative skill; these features are called *high-level interpretations*. This layer takes as input assessment of mid-level behavioral features and employs psychometric models to make inference about the competency of interest.

11.2.2 Using Multimodal Analytics to Study Influence of Affect and Noncognitive Behavior on Collaborative Study

Various studies have demonstrated the impact and influence of student affective state and behaviors such as turn taking (Woolley, Chabris, Pentland, Hashmi, & Malone, 2010), entrainment (convergence), and mirroring of affect (Lakin, Jefferis, Cheng, & Chartrand, 2003) on higher group intelligence and learning outcomes. Here a pilot study is presented that utilizes multimodal analytics to understand the incidence, dynamics, and influence of affect in collaborative problem solving (Luna Bazaldua et al., 2015). Our hypothesis is that performance on collaborative tasks is closely related to participant affective states and behaviors. Therefore, information about such states and behaviors can be important evidence for assessing the overall success of collaboration and individual ability to collaborate, as well as how well different tasks encourage collaboration. To test this approach a study was conducted involving 12 unique dyads collaborating in an online game-like science assessment, Tetralogue (Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2016; Zapata-Rivera et al., 2014). This platform includes both traditional assessment components, such as a set of multiple-choice items on general science topics, a simulation-based assessment, a personality test, and a set of background questionnaires. The simulation task relates to geology topics. The simulation-based task was developed as a task for individual test takers who will interact with two avatars, and as a collaborative task that requires collaboration among two human participants and two avatars in order to solve geology problems. The participants, who may be in different locations, interact through an online chat box and system help requests (i.e., opting to view educational videos on the subject matter).

Figure 11.2 illustrates the use of the Tetralogue collaborative activity platform and data capture system. Multimodal data, including video and activity log files, of each participating dyad were captured. The log files contain behavior that included frequency and content of chat messages between dyads, response to science questions both individually and as a dyad (jointly as a group), and system help requests (i.e., the participant asks to view educational videos on the subject matter to better answer assessment questions). The video data, on the other hand, recorded participant nonverbal behavior, which was analyzed on a frame-by-frame basis using automated facial expression classifiers and annotated by trained human raters on high-level noncognitive behaviors, including: affect display gestures, engagement, anxiety, and curiosity. The data were analyzed at individual and dyad levels and results derived using hierarchical clustering analysis demonstrated statistically significant evidence of cognitive and noncognitive behavioral convergence among dyads (see Sect. 3.2 for details).

11.2.3 Multimodal Data and Low-Level Features

Facial expression analysis of the video data was performed using the FACET SDK, a commercial version of the Computer Expression Recognition Toolbox (CERT; Littlewort et al., 2011). This tool recognizes fine-grained facial features, or facial action units (AUs), described in the Facial Action Coding System (Ekman, Friesen, & Hager, 1977). FACET detects human faces in a video frame, locates and tracks facial features, and uses support vector machine-based classifiers to output

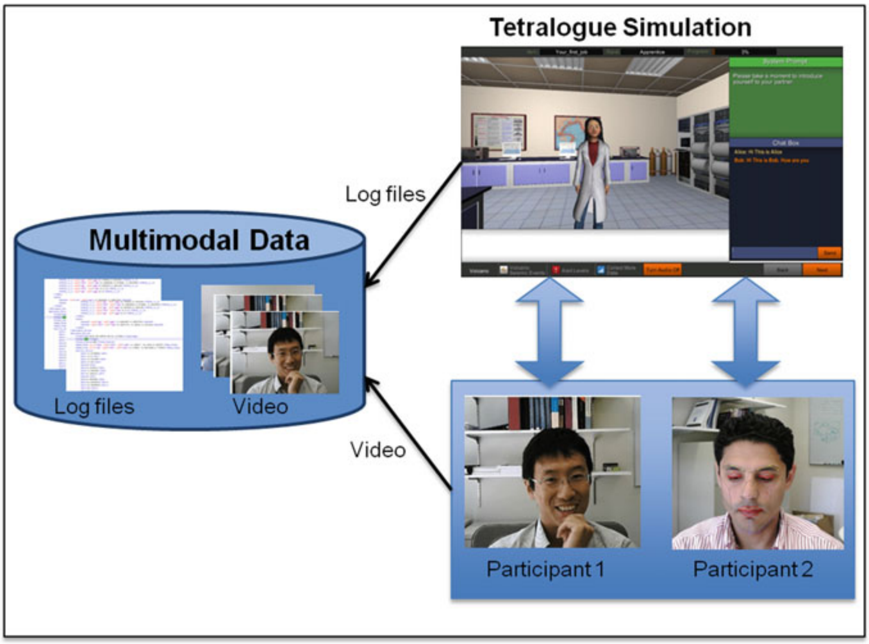


Fig. 11.2 Multimodal data capture including video and action log files while participants engage in collaborative activity on the Tetralogue platform

frame-by-frame detection probabilities of a set of facial expressions: anger, joy, contempt, and surprise.

In addition, seven trained coders reviewed and coded the videos using Anvil software (Kipp, 2001). The video data for each participant were assigned to two raters for annotation; however, in three cases there were three raters coding the same video file, and in two cases only a single rater was available for annotation. The raters followed the same coding scheme during the annotation process, that is, coding data with the labels: *hand touching face*, *expressing engagement*, *expressing anxiety*, or *expressing curiosity*. The behaviors were coded on a binary scale, reflecting whether the behaviors were absent or present. As an outcome of the annotation process, the Anvil software produced extensible markup language (XML) files that were parsed using the XML package (Lang, 2013) in R.

Engagement, anxiety, and curiosity were included in the annotation scheme because of the incidence and relevance with which these three noncognitive states occur in simulation games and online learning systems (Baker, D'Mello, Rodrigo, & Graesser, 2010; Calvo & D'Mello, 2010; D'Mello & Graesser, 2012; Woolf et al., 2009). The coding also included *hand touching face*, an affect display gesture that has been linked to affective and cognitive states such as boredom, engagement, and thinking (Mahmoud & Robinson, 2011; Whitehill et al., 2014).

11.2.4 Mid-level Features and Construct Inference

In order to study evidence of behavioral convergence, features from log files and video data of each of the 24 study participants were represented as a multidimensional behavioral feature vector composed of both the cognitive behaviors: *number_of_messages* and *number_of_help_requests* and the noncognitive behaviors (i.e., fraction of the time each participant exhibited the behavior): *engagement, hand_on_face, anxiety, curiosity, anger, joy, contempt,* and *surprise*.

An unsupervised, agglomerative hierarchical cluster analysis using an average linkage function was performed on a Euclidean distance matrix (i.e., a similarity matrix) computed from the multidimensional behavioral feature data of the study participants. Our hypothesis is that behavioral convergence will manifest in the cognitive and noncognitive features such that members of the same dyad will tend to group together from the beginning of the clustering process, that is, they will be closer to each other in the feature space than to others.

The similarity matrix of behavioral feature distances for participants within and outside dyad clusters was analyzed. Behavioral convergence would imply that, for dyad members, the average distances in feature space is smaller in a statistically significant manner than those of nominal dyad members. Moreover, to study the relative impact of cognitive and noncognitive features, two additional similarity matrices were computed: one using exclusively the cognitive features from log files (number of chat messages and number of system help requests) and the other using exclusively noncognitive features produced from the video data (the four facial expression detectors, and the four features from the coding scheme). All features were normalized to present equivalent scaled values between 0 and 1.

Table 11.1 shows the means and standard deviations of feature similarity distances of participants in dyad and nominal dyad populations. The results consistently show smaller average distances for the dyads (i.e., members within dyads displayed behavior that was more similar to each other than to others), supporting the convergence premise. Additionally, Student's *t* test was used to evaluate the statistical significance of these results. The results show that, when using both cognitive and noncognitive features together, the feature distance between participants belonging to the same dyad was smaller than the corresponding distance

Table 11.1 Average and standard deviation of behavioral feature distances within and outside dyads

Features	Populations	Mean	SD
Cognitive and noncognitive	Dyads	0.572	0.228
	Nominal dyads	0.730	0.243
Cognitive only	Dyads	0.365	0.216
	Nominal dyads	0.571	0.209
Noncognitive only	Dyads	0.411	0.178
	Nominal dyads	0.414	0.225

between nominal-dyads in a statistically significant manner: $t = 2.335$, $df = 11.7$, $p < 0.02$. However, when using noncognitive features alone, a statistically significant pattern of behavioral convergence was not found.

11.3 In Vivo Assessments of Nonverbal Behavior Using Multimodal Wearable Sensors

Human behavior modeling has been studied in a variety of disciplines such as behavioral science, social science, cognitive science, and artificial intelligence, among others. Several researchers have developed models of human behavior, from cognitive and affective states to human activities. This research has also explored the impact and influence of individual personality traits on outcomes of collective group activity. The traditional approach has been to create personality profiles using tools such as Big Five (Tosi, Mero, & Rizzo, 2000) or FACETS (Kyllonen, Lipnevich, Burrus, & Roberts, 2014) and analyzing the outcomes of the group interaction task vis-à-vis individual personality traits. Typically, this entails participants completing pretask or posttask questionnaires, an activity that is time intensive, expensive, and may induce subjective and social biases. Moreover, moment-by-moment activities and interactions in the group task are not captured, and the data are usually too sparse and coarse for an exploratory behavioral analysis. In contrast to this, some exciting new research has focused on measuring and modeling interpersonal behavior using low-level nonverbal behavioral data from environmental and wearable sensors (Olguin & Pentland, 2010). Of particular interest is research on assessing interpersonal skills in tasks such as negotiations, collaboration, leaderless tasks, and so forth, by tapping into a nonverbal, subconscious channel of human communication that Pentland calls honest signals (Pentland, 2008).

Multimodal analytics to conduct such assessments during in vivo group exercises (real world, in person) in a noninvasive manner using wearable sensors. These honest signals influence the outcome of group tasks, and therefore information about such states and behaviors can be important evidence for assessing the overall success of collaboration, individual ability, and interpersonal skills, as well as an alternative way to measure personality traits in and of themselves.

11.3.1 Analyzing Nonverbal Behavior

The subtle, unconscious patterns in which humans interact reveal their attitudes toward each other. These honest signals, as characterized by Pentland, are composed of patterns in physical activity, speech activity, and proximity, among other low-level behavioral cues. This research (Pentland, 2008; Woolley et al., 2010) has delineated a number of noncognitive, nonverbal behaviors that influence interpersonal interactions and will be the focus of this study. In particular we are interested in the following:

- Mimicry: The extent to which people in a conversation are reflexively mirroring each other.
- Conversational turn-taking: Participation balance and dominance.
- Activity: Measured as body movement or speech energy; increased activity often indicates interest and excitement.

One of the first attempts to measure face-to-face interactions between people using wearable sensors was the sociometer (Choudhury & Pentland, 2003). This wearable sensor package was used to learn social interactions from sensory data and model the structure and dynamics of social networks. Pentland described several statistical learning methods that use wearable sensor data to make reliable estimates of users' interactions. He presented a detailed description of behavior modeling for learning and classifying user behavior from proximity and location data, and influence modeling for predicting the behavior of a subject from another subject's data.

In an ongoing pilot study conducted at ETS, wearable sensors, specifically the Sociometric Badge (Olguin & Pentland, 2010) are utilized to measure nonverbal behavior in human interactions. The Sociometric Badge is a wearable sensing device that can be used to study human behavior and social interactions. Specifically, the badge collects information on (a) speech features such as volume, tone of voice, and speaking time; (b) body movement features such as energy and consistency; (c) information regarding people nearby wearing a Sociometric Badge; (d) the proximity of Bluetooth-enabled devices; and (e) approximate location information. The badges will not record speech or conversational content (unless this option is manually enabled). Figure 11.3 shows an image of the wearable badge. The study consisted of 24 participants that were divided into groups of four to work on a decision-making task. Each group member was given a role (e.g., Vice President [VP] of Finance, VP of Operations) of a hypothetical company, and the groups were tasked with choosing a store location that would be best for their company as it moves into a foreign market. Each member was given positive, negative, and neutral information about each of three potential store locations. Participants wore Sociometric Badges that recorded features from speech and body motion. Figure 11.4 illustrates preliminary analysis that easily shows evidence of turn taking and dominance of social interactivity. In the top left image, each participant is represented by a node (colored circle) in a connected graph. The thickness of the edges connecting any pair of nodes represents the number of speaking turns between the participant pair. It can be clearly seen that the individuals represented by the red, blue, and orange nodes had more turns between themselves than with the person represented by the green node. A similar picture emerges in the top right image, which shows a pie chart of individual speaking time, and the bottom image, which shows a timeline of speech onsets and offset from each of the four participants.

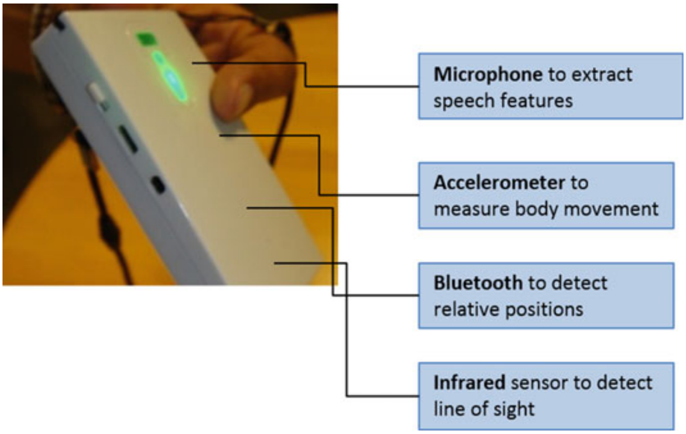

Microphone to extract speech features
Accelerometer to measure body movement
Bluetooth to detect relative positions
Infrared sensor to detect line of sight

Fig. 11.3 The wearable sociometric badge

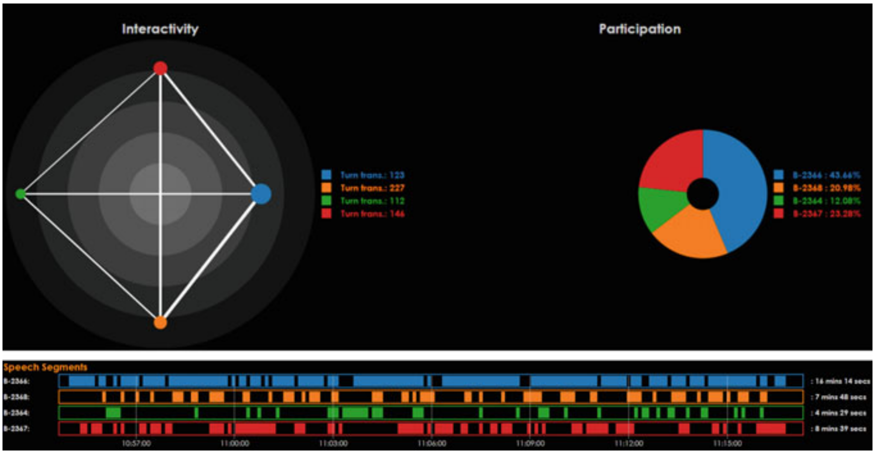


Fig. 11.4 Speech frequency and segments of a four-person group measured with sociometric badges

11.4 Concluding Remarks

This paper presents a framework to design and develop computational models that enable analysis of noisy, unstructured, multimodal data for the capture, analysis, and measurement of complex human behavior. This approach utilizes multimodal data including audio, video, and activity log files and constructs a hierarchical analysis methodology to model temporal dynamics of human behavior and the integration of multiple data modalities. The efficacy of such a methodology is demonstrated with two pilot studies where this approach was implemented to study collaborative learning and in vivo measurement of nonverbal behavior using wearable sensors.

Acknowledgements This research has benefited from help in data collection efforts, technical insight and valuable feedback the author received from colleagues including Diego Luna Bazaldua, Alina von Davier, Jiangang Hao, Robert Mislevy and Ketly Jean Pierre.

References

Amer, M. R., Siddiquie, B., Khan, S., Divakaran, A., & Sawhney, H. (2014). Multimodal fusion using dynamic hybrid models. In Institute of Electrical and Electronics Engineers (Ed.), *2014 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 556–563). Los Alamitos, CA: IEEE.

Baker, R. S., D’Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*(4), 223–241.

Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement, 21* (2), 175–189.

Calvo, R. A., & D’Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18–37.

Choudhury, T., & Pentland, A. (2003). Sensing and modeling human networks using the sociometer. In Institute of Electrical and Electronics Engineers (Ed.), *Proceedings of the 7th IEEE International Symposium on Wearable Computers* (pp. 216–222). Los Alamitos, CA: IEEE.

D’Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157.

Ekman, P., & Friesen, W. V. (1977). *Facial action coding system.*

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., et al. (2012). *Teaching adolescents to become learners: The role of noncognitive factors in shaping school performance: A critical literature review*. Chicago, IL: Consortium on Chicago School Research.

Fisch, S. M. (2005). Making educational computer games “educational.” In Association for Computing Machinery (Ed.), *Proceedings of the 4th International Conference for Interaction Design and Children* (pp. 56–61). New York, NY: Association for Computing Machinery.

Kipp, M. (2001) Anvil—A generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)* (pp. 1367–1370).

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill based, and affective learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78,* 311–328.

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., & Roberts, R. D. (2014). *Personality, motivation, and college readiness: A prospectus for assessment and development* (Research Report No. RR-14-06). Princeton, NJ: Educational Testing Service.

Lakin, J. L., Jefferis, V. E., Cheng, C. M., & Chartrand, T. L. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior, 27*(3), 145–162.

Lang, D. T. (2013). *XML: Tools for parsing and generating XML within R and S-Plus* (R package version 3.98–1.1). Retrieved from http://CRAN.R-project.org/package=XML

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J. R., et al. (2011). The computer expression recognition toolbox (CERT). In Institute of Electrical and Electronics Engineers (Ed.), *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 298–305). Los Alamitos, CA: IEEE.

Liu, L., Hao, J., von Davier, A., Kyllonen, P., & Zapata-Rivera, D. (2016). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on computational tools for real-world skill development*. Hershey, PA: IGI-Global.

Luna Bazaldua, D. A., Hao, J., Khan, S., Liu, L., von Davier, A. A., & Wang, Z. (2015). On convergence of cognitive and non-cognitive behavior in collaborative activity. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, … M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining Conference* (pp. 496–499). Madrid, Spain: International Educational Data Mining Society.

Mahmoud, M., & Robinson, P. (2011). Interpreting hand-over-face gestures. In S. D'Mello, A. Graesser, B. Schuller, & J-C Martin (Eds.), *Proceedings of the International Conference on Affective Computing and Intelligent Interaction* (pp. 248–255). New York, NY: Springer.

Morency, L. P., de Kok, I., & Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems, 20*(1), 70–84.

National Research Council. (2011). *Assessing 21st century skills*. Washington, DC: National Academies Press.

Olguin, D. O., & Pentland, A. (2010). Assessing group performance from collective behavior. In Association for Computing Machinery (Ed.), *CSCW 2010 Workshop on Collective Intelligence in Organizations*. New York, NY: Association for Computing Machinery.

Pentland, A. (2008). *Honest signals: How they shape our world*. Cambridge, MA: MIT Press.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 35,* 135–146.

Siddiquie, B., Khan, S., Divakaran, A., & Sawhney, H. (2013, July). Affect Analysis in natural human interaction using joint hidden conditional random fields. In Institute of Electrical and Electronics Engineers (Ed.), *2013 IEEE International Conference on Multimedia and Expo (ICME 2013)* (pp. 1–6). Los Alamitos, CA: IEEE.

Smarter Balanced Assessment Consortium. (n.d.). *Thermometer crickets: Grade 11 performance task*. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/performance-tasks/crickets.pdf

Tosi, H. L., Mero, N. P., & Rizzo, J. R. (2000). *Managing organizational behavior* (4th ed.). Cambridge, MA: Blackwell Publishers.

Whitehill, J., Serpell, Z., Lin, Y. C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing, 5*(1), 86–98.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: recognizing and responding to student affect. *International Journal of Learning Technology, 4*(3–4), 129–164.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686–688.

Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., & Katz, I. R. (2014). *Assessing science inquiry skills using trialogues. Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 8474, pp. 625–626). Berlin, Germany: Springer.

Chapter 12
Measuring Collaboration in Cross-Cultural Contexts

C. Shawn Burke, Jennifer Feitosa, Eduardo Salas, and Michele Gelfand

Abstract The use of team-based organizing has become the norm in many organizations, especially those characterized by complexity. However, research has shown that teams are not always effective, despite their popularity (Sims & Salas, 2007). The complex endeavor of creating and maintaining the enabling conditions for team performance is further compounded when teams are culturally diverse. While cultural diversity can provide synergies, research has shown that it can also lead to process loss as members attempt to navigate differences in attitudes, beliefs, and values that often remain hidden under the surface and impact team interaction. Therefore, the purpose of this chapter is twofold. First, to highlight some of the areas that may prove challenging for culturally diverse teams to navigate; and, in turn, provide insight into what constructs to measure. Second, to identify a set of measurement considerations that must be navigated when assessing collaboration within culturally diverse teams.

Keywords Cultural diversity · Multicultural teams · Cross-cultural teams · Measurement · Assessment · Collaboration

C. Shawn Burke (✉)
University of Central Florida, Orlando, USA
e-mail: sburke@ist.ucf.edu

J. Feitosa
City University of New York, New York, USA
e-mail: jennifer.feitosa81@brooklyn.cuny.edu

E. Salas
Rice University, Houston, USA
e-mail: Eduardo.Salas@rice.edu

M. Gelfand
University of Maryland, College Park, USA
e-mail: mgelfand@umd.edu

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_12

12.1 Introduction

There was a time when the use of team-based organizing was a competitive advantage for organizations; now however, the use of teams is the norm in many organizations. Moreover, research has shown that teams are not always effective despite their popularity (Sims & Salas, 2007). When faced with a complex task, organizations often pull together a team of experts to collaboratively work on the task. What is often not recognized by the organization is that a team of experts does not equal an expert team. Effective taskwork is a necessary, but not sufficient condition for effective team performance—members must also be able to effectively coordinate among themselves to accomplish interdependent tasks (Morgan, Glickman, Woodard, Blaiwes, & Salas, 1986). While creating and maintaining the enabling conditions for team performance is a complex endeavor, the complexity is further compounded within culturally diverse teams. While there are many ways in which teams may be culturally diverse (e.g., nationally, organizationally, professionally), for the purposes of this chapter we focus on cultural diversity with respect to national culture as evidenced by the "shared norms, values, and practices of a nation" (Helmreich, 2000, p. 134). Newman and Nollen (1996) defined national culture as, "the values, beliefs, and assumptions learned in early childhood that distinguish one group of people from another" (p. 754). Similarly, Sato (2016) defined culture as "the values, beliefs, and norms that influence the way we understand, engage in our experiences, and respond to the range of situations we face" (p. 4). Following from this, culturally diverse teams can be defined as those teams whose members possess differences in values, beliefs, and preferences for cognition and action that are driven by the national culture with whom they identify.

While cultural diversity can provide synergies for the team, research has shown that it can also lead to process loss as members attempt to navigate differences in attitudes, beliefs, and values that often remain hidden under the surface and impact team interaction. For example, Klein (2004) stated, "Intelligent and thoughtful people from different national groups sometimes identify different problems, make different plans, negotiate and coordinate differently, and make different decisions during complex cognitive tasks" (p. 250). Research has shown that cultural differences influence communication, group cohesiveness (Arman & Adair, 2012), trust (Brown, Adams, Famewo, & Karthaus, 2008), attitudes (Boyd et al., 2009), and moods (Kanas et al., 2009). However, when teams can navigate these differences, research has shown that such teams can outperform homogeneous teams in the long term. So the question becomes, how can we assist team members in uncovering the landmines that remain hidden? One path forward is by understanding where potential landmines are and assessing teams around these areas, such that they can work through differences to create a hybrid culture that can drive team performance. Therefore, the purpose of this chapter is twofold. First, we will begin to highlight some of the areas that may prove challenging for culturally diverse teams to navigate. This, in turn, not only provides insight into what constructs on which to initially focus measurement, but also highlights how, within culturally diverse teams, the notion of what is “effective” may differ. Our second purpose is to identify a set of challenges that practitioners may face in assessing collaboration within culturally diverse teams.

12.2 What Should Be Measured? Identifying Possible Break Points

Team process has been defined as, “members’ interdependent acts that convert inputs to outcomes through cognitive, verbal, and behavioral activities directed toward organizing taskwork to achieve collective goals” (Marks, Mathieu, & Zaccaro, 2001, p. 357). Recent work has introduced the notion that teams perform in “temporal cycles of goal-directed activity, called *episodes* (Marks et al., 2001, p. 359). Embedded within these episodes are action and transition phases. Action phases are those “periods of time when teams are engaged in acts that contribute directly to goal accomplishment (i.e., taskwork);” whereas, transition phases are those “periods of time when teams focus primarily on evaluation and/or planning activities to guide their accomplishment of a team goal or objective” (p. 359). Throughout the remainder of this chapter, the Marks et al. (2001) framework is used as a way to begin to highlight the places where cultural diversity may pose challenges for teams and, thereby, the types of processes that are important to assess. This framework was chosen because it represents current thinking in the teams literature, and meta-analytic work has shown the predictive power of the included processes with respect to team performance and team satisfaction (LePine, Piccolo, Jackson, Mathiue, & Saul, 2008). Due to space constraints, rather than focus on all the team processes within each phase (see Table 12.1) we highlight a sampling of processes where cultural diversity could have the largest impact.

12.2.1 Transition Phase Processes

12.2.1.1 Mission Analysis

There has been little work explicitly conducted on the impact of cultural diversity within teams and mission analysis; however, leveraging work on cultural orientations and values with that on mission analysis can provide a basis to specify potential impacts (see Table 12.1). Power distance refers to the “extent to which a

Table 12.1 Taxonomy of team process (adapted from Marks et al., 2001)

Phase definition	Team process	Defined	Key components
Transition phase "Periods of time when teams focus primarily on evaluation and/or planning activities to guide their accomplishment of a team goal or objective" (Marks et al., 2001, p. 364)	Mission analysis	"Interpretation and evaluation of the team's mission, including identification of its main tasks as well as the operative environmental conditions and team resources available for mission execution" (Marks et al., 2001, p. 365)	Development of shared mental models Backward evaluation Forward visioning
	Goal specification	"The identification and prioritization of goals and subgoals for mission accomplishment" (Marks et al., 2001, p. 365)	Identification and a clear articulation of goals Goals aligned with strategies How to achieving goals Expectations Relay of task-related information Prioritization Role assignment Communication of plans
	Strategy formulation	Strategy formulation and planning involves developing additional options of methods to complete the mission	
Action phase Teams conduct activities that have been identified as leading to the accomplishment of goals	Monitoring progress	The ability to accurately monitor and assess the situation	Assess current state Identify unique aspects of the teams
	System monitoring	Involves assessing and monitoring the resources of the team, as well as monitoring the external environmental factors that could relate to outcomes and successes of the mission	Assessing and monitoring resources Monitoring external environmental factors
	Monitoring and backup behaviors	"Assisting team members to perform their tasks, which may occur by (1) providing verbal feedback or coaching, (2) assisting a teammate behaviorally in carrying out actions, or (3) assuming and competing a task for a teammate" (Marks et al., 2001, p. 367)	Team members assisting each other Team members seeking assistance Role clarity
	Coordination	"Process of orchestrating the sequence and timing of interdependent actions" (Marks et al., 2001, p. 363)	Communication between team members Team member adaptability Effective planning Plan of action adjustment

(continued)

Table 12.1 (continued)

Phase definition	Team process	Defined	Key components
Interpersonal processes "Processes teams use to manage interpersonal relationships" (Marks et al., 2001, p. 368)	Conflict management	Process of handling conflict so that it does not negatively affect team performance. Consists of preemptive and reactive conflict management. "Preemptive conflict management involves establishing conditions to prevent, control, or guide team conflict before it occurs. Reactive conflict management involves working through task and interpersonal disagreements among team members" (Marks et al., 2001, p. 363)	Identify conflict within the team Problem-solving Compromise Cooperation Establish norms for handling conflict
	Motivation and confidence building	"Generating and preserving a sense of collective confidence, motivation, and task-based cohesion with regard to mission accomplishment" (Marks et al., 2001, p. 363)	Encourage team members to achieve and maintain high performance Exhibit confidence in the team's ability to successfully accomplish its task/mission
	Affect management	Regulating team member emotions during and after taskwork	Emotion regulation Boost morale and cohesion

society accepts the fact that power in institutions is distributed unequally" (Hofstede, 1980, p. 45). Team members who value high power distance would expect mission analysis to be conducted primarily by the team leader, and subordinates would not feel comfortable speaking up to offer input. This is in direct contrast to the expectations of those who value low power distance, where input by all would be expected. Furthermore, if it is the team leader who values high power distance, subordinate input may not be elicited. All of these dynamics combine to result in a situation whereby information exchange may be limited and therefore hinder mission analysis.

Cultural differences in analytic/holistic reasoning may also pose challenges to mission analysis. Specifically, cultural differences in reasoning styles may result in different cues being sampled as well the processes engaged in during decision making (Choi & Nisbett, 2000). An analytic orientation is characterized by the use of logic and dispositional cues, whereas holistic reasoning relies heavily on dialectical reasoning, with little use of formal logic. Other cultural dimensions that have been argued to impact mission analysis include: uncertainty avoidance, high-low context, and field dependence (Salas, Burke, Wilson-Donnelly, & Burke, 2004). For example, those less comfortable with uncertainty may not be willing to step out of their "comfort zone" and thereby may not sample the same types of cues as those more comfortable with uncertainty (Salas et al., 2004). Preference for high or low context will impact how meaning is derived from verbal or written communication, with low context cultures focusing primarily on what is explicitly stated; in contrast, high context cultures derive meaning based on the surrounding context of the message and implicit cues (Ting-Toomey, 1999). Finally, cultural preferences for past or future orientation (see Hall & Hall, 1990) may cause individuals to place differential weight on the information that is collected; thereby, impacting mission analysis.

12.2.1.2 Strategy Formulation

There are many ways in which cultural diversity on a team may impact strategy formulation and therefore illustrate the instrumentality in assessing it (e.g., tolerance for uncertainty, power distance, hypothetical/concrete reasoning). Team members whose cultural orientation is characterized by a low tolerance for uncertainty find uncertainty stressful. With respect to strategy formulation, these members would be expected to (a) adhere strictly to rules and process (Hall & Hall, 1990); (b) be less comfortable making decisions with incomplete information (Hall & Hall, 1990); (c) value consensus (Lane & DiStefano, 1992); and (d) may feel unsettled during the decision making process and more reluctant to change a decision once fianlized (Helmreich & Merritt, 1998). Variations on this dimension may lead to conflict (Klein, 2004) regarding the degree of plan specificity needed, norms for challenging status quo, and plan reformulation/adaptation.

Culture may also impact comfort with contingency plans and mental simulations during strategy formulation. For example, differences in hypothetical and concrete reasoning might impact the degree to which "what-if-ing" is valued (Klein, 2004). This not only impacts strategy formulation, but the team's ability to be adaptive as what-if-ing and mental simulations increase the breadth of members' mental models; thereby, guiding future action.

12.2.2 Action Phase Processes

12.2.2.1 Monitoring Progress Toward Goals

There has been limited work conducted here with respect to culture; however, cultural orientations that view time as a scarce resource may engage in more monitoring of goal progress (Arman & Adair, 2012). Conversely, team members from cultures that have a low tolerance for uncertainty may be expected to monitor goal progress more frequently than team members from cultures with a high tolerance for uncertainty. Greater frequency in monitoring goal progress serves to facilitate situation awareness and thereby avoids uncertainty to the extent the environment allows.

12.2.2.2 Team Monitoring and Backup Behavior

Unlike the case with many of the other team processes, research has begun to explicitly examine the relationship between cultural values and team monitoring/backup behavior. In thinking about team monitoring/back-up behavior there is inherently an evaluation and feedback (even if implicit) component. Related to this, collectivists have been found to evaluate in-group members more positively than do individualists (Gomez, Kirkman, & Shapiro, 2000). Other research, while not explicitly focusing on teams, does offer additional insight with respect to the feedback or assistance portion of monitoring/back-up behavior. Research has shown that cultures that adopt a root cause orientation are more likely to attribute responsibility to the person than those with a systems orientation where attributions are more context based (Schweder & Bourne, 1982). This, in turn, may impact team members' perceptions of one another's ability. These attributional differences may also impact the degree to which feedback is sought out or accepted. Members from cultures with a root cause orientation would be more likely to seek out feedback as they view it as part of the learning/improvement process (Klein, 2004). This is in direct contrast to those with a systems orientation where feedback may be seen as an attack on the individual as compared to an appraisal of a specific capacity.

Cultural orientations regarding power distance may have an impact on team backup behavior. Members with an orientation indicative of low levels of power distance will be more willing to accept and offer verbal input and assistance without consideration of team member status. Conversely, members from high power distance cultures would not be expected to be comfortable seeking assistance from those members of lower status, nor would low status members be expected to provide monitoring or backup behavior to team leaders (Klein, Klein, & Mumaw, 2001).

12.2.2.3 Coordination

While there has been limited work that directly investigates the manner in which cultural diversity may impact coordination, most of the cultural differences discussed up to this point could ultimately be argued to impact the team's ability to coordinate. One of the mechanisms that has been consistently argued to allow teams to coordinate is the presence of shared mental models (Langan-Fox, Anglim, & Wilson, 2004; Rentsch & Woehr, 2004). Cultural differences in values, attitudes, beliefs, and preferences reflected in one's cultural orientation often cause the underlying cognition (which forms mental models and shared mental models) among team members to vary. These differences are often not seen on the surface; thereby, team members' make implicit assumptions about how other team members view the team or team processes which are often incorrect. These assumptions guide member's mental models and corresponding behavior, and within multicultural teams they often lead to lower levels of shared mental models, which in turn, impact coordination.

12.2.2.4 Communication

While communication is not one of the Marks et al. (2001) dimensions, communication is one of the most researched aspects of cultural diversity in teams. Researchers have shown that cultural diversity often leads to miscommunication (Adler, 1997; Humes & Reilly, 2007; Kealey, 2004; Li, 1999). Cultures not only have different languages which can lead to information being lost, but different norms for communication also exist which can cause major problems in culturally diverse teams (Berger, 1996). Research onboard the International Space Station (ISS) found differences in direct and indirect communication, individual recognition preferences, and comfort with participation in large group conversations (David, Rubino, Keeton, Miller, & Patterson, 2011). The researchers found instances where "high power distance and collectivist Russian crew members are apprehensive about participating in large group conversations" (David et al., 2011, p. 11).This, in turn, can limit the sharing of information and diverse perspectives.

Cultural differences will also impact information exchange. Information exchange has been argued to be slower in multicultural teams as increased effort is needed to calibrate meaning (Cherrie, 1997; Helmreich, 2000). Conyne, Wilson, Tang, and Shi (1999) found that individualists were less likely to direct communication to the leader as compared to collectivists. While collectivists were more hesitant to provide information, when they did speak they did so for longer periods. Communication style has also been found to differ across cultures. Specifically, individualists show a preference for direct communication, while collectivists prefer an indirect style (Gudykunst et al., 1996). These differences in information exchange could impact both the sharing of relevant information as well as ensuring that information is transmitted to the right person.

12.2.2.5 Leadership

Team leadership has been argued to be able to make or break a team. While limited, the research that has examined team leadership in multicultural teams has described it as challenging (Oertig & Buergi, 2006). The potential for misunderstandings, miscommunication, and conflict within intercultural teams leads to dynamics which are often complex and time consuming (Humes & Reilly, 2007). Interpersonal leadership has been reported as being emphasized within the teams to facilitate working through the different perspectives often present (Watson, Johnson, & Zgourides, 2002). Once interpersonal issues were resolved, culturally diverse teams were able to perform more effectively than culturally homogeneous teams on a problem-solving task

Prior research has indicated that cross-cultural differences related to team leadership may alsohave an impact. For example, the proficiency and comfort with which individuals can manage relationships and network have been shown to be key leadership behaviors and vary across cultures (House, Hanges, Javidan, Dorfman, & Gupta, 2004). Additionally, cultures vary in their preference for doing versus being (Kirkman & Shapiro, 2001); these variations can impact the leadership process. Individuals from cultures with a doing orientation tend to value action as compared to those with a being orientation which values reflection and understanding (Kluckhohn & Strodbeck, 1961). These differences will not only drive relationship management, but mission analysis, planning, goal-setting, and strategy formulation processes. Also related are findings indicating Chinese leaders reported valuing group success over individual member's feelings about participation (Conyne et al., 1999).

12.3 Measuring Key Processes in Culturally Diverse Teams

The above section on potential breakpoints within culturally diverse teams highlights points where measurement needs to focus such that issues can be identified early, before diverse teams become entrained into a rhythm where team coherence (i.e., shared affect, behavior, and cognition which drives team performance) is less than optimal. The longer decrements in team coherence remain undiagnosed, the greater the propensity for faulty mental models to form and negative affect to take hold, thereby limiting information sharing and potentially resulting in cliques representing in- and out-groups within the team. Next, we highlight some challenges that practitioners (and researchers) may face in measuring team process within culturally diverse teams, as well as a few potential ways forward. Most of the identified challenges cut across the processes described earlier, but when relevant, particular examples will be elaborated upon.

Team performance measurement is not new, but it is often one of the most overlooked and misunderstood components within team development, and very little of the literature explicitly focuses on the challenges of measuring culturally diverse teams. Pulling from the literature on team performance measurement, several criteria can be extracted in regards to the characteristics of sound team performance measurement. For example, researchers have argued that team performance measures should (a) be competency based or theory driven, (b) be contextualized and task relevant, (c) collect from multiple sources, (d) be descriptive of team performance, (e) capture the dynamic and longitudinal nature of team performance, and (f) capture and discriminate between multiple levels of performance (Rosen et al., 2012). However, little of this work speaks to how these things may manifest themselves in designing measures for culturally diverse teams. Gelfand, Raver, and Ehrhart (2002) identified several methodological criteria that must be met with regard to cross-cultural research, including (a) the use of well-developed theories to guide the development of research questions, measures, and the sampling of cultures; (b) employment of methods to guard against the use of etic constructs, and (c) ensuring that the methodology that is chosen is culturally appropriate. Integrating the requirements for team performance measures with those of cross-cultural research, we delineate how these requirements translate into considerations for those charged with developing such measures and finding a way to move forward.

12.3.1 Consideration 1: What to Measure/What Guidance Is Available?

Traditionally, the determination of what knowledge, skills, attitudes, and abilities (KSAOs) to measure would be guided by a job and/or task analysis (see Brannick, Levine, & Morgeson, 2007 for further information) and in the case of teams, a team task analysis (see Burke, 2005 for further information). This process is augmented by theoretical and empirical frameworks that delineate those team processes and emergent states that have been shown to drive effective team performance (see Marks et al., 2001; Mathieu, Maynard, Rapp, & Gilson, 2008; Salas, Shuffler, Thayer, Bedwell, & Lazzara, 2015). This is no different when building measures for culturally diverse teams; however, the processes that often comprise these activities must be filtered through a wider cultural lens that takes into account how differences in cultural orientation (e.g., values, beliefs, norms, attitudes, preferences for action) might impact how the processes are implemented. For example, while there has been a tremendous amount of work investigating the drivers (e.g., antecedents, processes, states) of effective team performance, much less work has been conducted that truly examines those drivers in the context of culturally diverse teams. Within much of the current literature, comparisons are made between homogeneous groups, providing a picture of interactions within teams of different cultures (if results are compared across teams). For example, Conyne et al. (1999) examined discussion group member interactions within American and Chinese teams. In Chinese teams, members were more likely to direct communications to the leader versus to other group members, while the opposite was found in American teams. There were also differences between Chinese and American team members in leadership approaches, team member communication styles, and team members' willingness to speak up.

While this research is instructive for culturally homogeneous teams, it offers little insight into how Chinese and American team members are likely to interact within the same team. Studying only teams that have within-team homogeneity fails to provide information about how culturally distant values and preferences interact within a team setting, thereby leaving a gap in the literature and a corresponding lack of guidance for those responsible for training and assessment of culturally diverse teams. The information set forth earlier in this chapter provides a starting point in terms of highlighting points at which team process may be most likely to break when operating within culturally diverse teams (and thereby pointing to an important area to measure), but does little to talk about how the actual cultural composition of the team might impact team process. More research is needed in this area to better guide those in charge of developing measures.

When we develop measures for culturally diverse teams, the need to look at the processes of team measurement development through a multifaceted cultural lens becomes important, not only in thinking about how the team processes might interact, but also in thinking about how the actual team task analysis might be implemented (see Arthur, Villado, & Bennett, 2012; Burke, 2005). While team task analysis seeks to identify those tasks where coordination demands are present and determine what KSAOs are involved, many of the methodologies used are similar to those in traditional job and task analyses (i.e., interviews, surveys, observations, analysis of archival documents). Therefore, when dealing with culturally diverse teams, we must consider how team members' culture might impact how they respond to the survey, how they respond to the person who is doing the interview, and the degree to which focus groups might be acceptable. Many of these considerations are covered in later portions of the chapter, so they will not be further detailed here.

12.3.2 Consideration 2: Did I Guard Against Use of Etic Constructs?

One of the primary challenges in research on culturally diverse teams is the need to ensure that measures are capturing the emic view (i.e., the culture-specific meaning of a construct; the insider view of a construct) as opposed to the more commonly seen etic view (i.e., an outsider's perspective on what a particular construct means). This state of affairs exists because the majority of the work that has been conducted on teams, investigating the drivers of effective team performance, their manifestation, and their assessment, has been done in the West. This poses a challenge for those wanting to assess team performance within culturally diverse teams, for research suggests that cultures vary in their mental models about how teams should operate, as indicated by the types of metaphors they use in describing teams. This, in turn, has implications for how members conceptualize team processes within the team. For example, Gibson and Zellmer-Bruhn (2001, 2002) found five teamwork metaphors that emerged across cultures (i.e., family, sports, community, associates, and military). These metaphors provide insight into how individuals from different cultures view the scope, norms, and breadth of teams, thereby beginning to provide an emic view of teams. Cultures which used a sports metaphor to talk about teamwork, described teams as having: a narrow scope where activity was limited to physical and social interaction, clearly defined roles, low levels of hierarchy, and specific and measurable objectives with clear consequences. Conversely, the military metaphor referred to teams as having a "fairly limited scope, with activity limited to professional, physical and educational activities" (Gibson & Zellmer-Bruhn, 2002, p. 8). Expectations include strong hierarchical roles and clear objectives.

Gibson and Zellmer-Bruhn (2002) also investigated the degree to which national culture predicted the use of specific teamwork metaphors. Results found that individuals from individualistic cultures used metaphors that were narrower in scope (i.e., sports, associate), while individuals from collectivist cultures tended to use metaphors indicative of teams being broader in scope (i.e., family, community). In addition, high levels of power distance was found to be related to the use of metaphors involving clear objectives (i.e., military). These metaphors, in turn, reflect preferences and expectations regarding how many of the team interaction processes described earlier in the paper may be operationalized—differences in these mental models (which lay under the surface) are one of the reasons some of these processes pose challenges in culturally diverse teams.

One of the few team processes that has been heavily examined across cultures is leadership. This work, in turn, may begin to provide insight into emic views of leadership. Specifically, researchers within the Global Leadership and Organizational Behavior Effectiveness (GLOBE) project have reported that implicit theories of leadership vary across cultures. Results suggest that universally endorsed styles include charismatic/value-based and team-orientated leadership; however, human and participative styles are near universally endorsed. In contrast, autonomous and self-protective dimensions are culturally contingent. High within-culture agreement was also found with respect to the implicit leadership styles (see also Brodbeck et al., 2000). As a whole, this work provides further evidence to earlier findings suggesting that cultures vary in their prototypes of effective leadership (Bass, 1997; Hofstede, 1993; Triandis, 1993). While being universally endorsed, charismatic group leadership was found to be more prevalent in collectivist cultures (Pillai & Meindl, 1998). Finally, across cultures transformational leadership was found to be correlated with group potency and member self-efficacy (Jung & Yammarino, 2001).

The work on team metaphors and leadership across cultures begins to provide some insight into emic views on these areas, yet the predominant number of team processes have not been investigated in this manner. Ultimately, this puts more burden on those doing the assessing to ensure that the construct they are assessing is really meaningful to the target and that the manner in which it is operationalized within the assessment is culturally valid. Cultural observations, interviews with subject matter experts in the region of interest, analysis of archival text, and the use of semantics can be tools that those charged with assessment can begin to use to ensure that emic constructs are being captured.

12.3.3 Consideration 3: Are My Methods Culturally Appropriate?

While there are a variety of methods available to those charged with assessing teams, the field is dominated by the use of surveys. This is true within both the teams and cross-cultural literature. Surveys have the advantage of being fairly simple to administer; however, within culturally diverse teams there may be differences among cultures for how instructions and terms are understood, motivation for completion, and response sets (Gelfand et al., 2002; Triandis, 1993). For example, Moshinsky (2000) found that for Russian participants, instructions to complete a survey independently were inconsistent with their cultural expectations and values, and therefore the Russians completed the task in more of a consensus format, which was closer to their mental model of how they should respond (as cited in Gelfand et al., 2002).

With respect to surveys, the manner in which they are worded or the scale anchors might also need to be taken into account when dealing with culturally diverse teams. For example, in more collectivistic cultures, having anchors that are clearly at opposite ends of the spectrum, whereby one anchor explicitly represents an ineffective team member, may be less effective due to members' reluctance to criticize members of their in-group. Similar dynamics may be seen in cultures with high power distance, in cases where the survey is asking a lower ranking member to report on someone with more status. Less variability of responses may be seen in both of these instances due to cultural norms. One potential way around this (other than being very cognizant of culturally-based power dynamics) is to devise scales such that the scale endpoints are designed to limit social desirability bias (in this case, with regard to culture) in that endpoints are framed with cultural preferences in mind (e.g., reframing negatively valenced items).

While surveys are definitely the most widely used methodology within teams research, other cross-cultural research methodologies that could be leveraged include knowledge elicitation techniques (see Cooke, Salas, Cannon-Bowers, & Stout, 2000; DeChurch & Mesmer-Magnus, 2010), direct observation, analysis of cultural artifacts (e.g., newspaper articles), and experimentation. With regards to those methodologies that involve another person serving as the data extractor (e.g., some forms of knowledge elicitation and observation) one must consider how cultural values and preferences for action might impact the relational dynamics (and therefore the quality of the responses). For example, interviews or think-aloud protocols are methods that have been used in knowledge elicitation. In these cases, cultural dynamics such as power distance, high-low context, and individualism/ collectivism may impact the use of this method, specifically impacting the comfort with which information is shared along with the verbosity of the response.

While challenging, this hurdle and may be one of the more easy ones to surmount. The lesson is that in choosing the methods by which to obtain the data you need to consider the cultural lens from which the instructions and format will be viewed. There are a variety of methods that are available; practitioners just need to have a well-developed methodological toolbox and think about the cultural context within which the assessment will be taking place.

12.3.4 Consideration 4: Are My Elicitation Sources Culturally Appropriate?

Closely related to the method is the elicitation source. While elicitation source (i.e., who is providing the information about the team) is always a consideration in all team measurement, it becomes a bit more complex in culturally diverse teams. Traditionally, the elicitation source can range from the individual with whom we are concerned (i.e., self-report) to a third party or parties (i.e., supervisor, leader, rater/observer) who are directly involved in measuring team performance. There are also a few rare instances where the elicitation source is the entire team. This, in turn, results in a consensus rating as compared to the more commonly obtained individual level ratings, which are later combined to form some type of team aggregate score. Good science would recommend that the choice of which elicitation sources are acceptable is driven by theory related to the construct one is assessing; however (e.g., who is in the best position to observe or report on the construct–behaviors more easily observed by an outsider versus attitudes or cognition), often the choice is driven by convenience. Within culturally diverse teams, this may be even more challenging as often team members are distributed over time and space. Additionally there are considerations regarding the appropriateness/acceptability of the data based on members' cultural expectations.

12.3.4.1 Self-report

Within the literature on team process measurement, perhaps the most common elicitation source is the team members themselves. The use of self-report as an elicitation source can be especially helpful in measuring perceptions of how well the team is interacting and working towards the team's goal (i.e., teamwork). Because the information comes directly from the individual completing the task, responses are subject to several traditional rater biases (as are all subjective measures). Perhaps the most common biases in self-reports are inflated scores. However, in culturally diverse teams, factors such as individualism/collectivism and power distance may impact self-report data. For example, collectivists are apt to be much more concerned about "saving face" for themselves and their in-group and therefore might be less likely to provide negative ratings on self-report scales. Differences in power distance and cultural views concerning gender roles may also impact the motivation and effort that participants put into completing self-report measures.

12.3.4.2 Supervisor Ratings

A second common source from which to collect data is the supervisor or team leader. The use of this data source provides a different perspective than self or peer ratings and may be best used for behaviors that are readily observable as compared to those that are more implicit. Supervisor ratings may bypass variance with respect to the individual members of the team (e.g., individual differences) and the manner in which the individual contributions combine to create the process as seen at the team or unit level. Some team nuances are often lost or at the very least not readily apparent as the supervisor applies his/her own weighting scheme to individual actions within the team. Within culturally diverse teams, members may weigh the value they place on these ratings differentially depending on leadership prototypes and cultural variations in gender roles.

12.3.4.3 Peer Ratings

An alternative to self or supervisor ratings is to use peer ratings. By enlisting the evaluation of peers (other team members), the evaluation of team processes may be closer to actual processes occurring within the team. Peer ratings of the degree to which individual team members engage in particular team behaviors is an increasingly common practice used to look at the structural aspects of team process and what the team networks are (this has been most commonly done when examining leadership; see Carson, Tesluk, & Marrone, 2007). Within culturally diverse teams, peer ratings may serve to mitigate some of the potential bias in self-ratings; however, peer ratings might also show greater variability in culturally diverse teams, based on cultural expectations and implicit theories regarding team interaction. It is important to note that evaluations through peer reports, as well as supervisor or self-reports, do not need to be used independently of one another to develop a picture of team processes. In actuality, the use of multiple sources in obtaining data is recommended, as it provides a fuller picture of the complexity of team functioning. The optimal sources (e.g., peer, self, supervisor) to use will vary based on the cultural composition of the team and the construct being assessed (i.e., how easily observable it is by others).

12.3.5 Consideration 5: When to Measure Team Process?

Temporal considerations in teams should be taken into account in determining not only the most efficacious time to measure, but also the content of the measurement. Several streams of research point to the importance of considering the temporal aspects of teams. For example, Marks et al. (2001) argued that teams perform in "temporal cycles of goal-directed activity, called *episodes*" (p. 359). These episodes consist of action and transition periods whereby the primary focus of the team differs depending on whether they are in an action or a transition period. In essence, action periods are those periods when the team is primarily focused on task accomplishment and is directly engaged in the task. In contrast, transition periods are those periods when the team's primary focus is on planning and regulatory activities that serve an evaluator function which, in turn, contributes as input to later action phases (Marks et al., 2001). Following from this is a recommendation that team process should be measured at the conclusion of an action phase so that feedback can be provided as input into the transition phase. Gersick (1988) offered insight into a second point at which it makes sense to assess team performance—at its midpoint transition. Gersick found that upon formation, teams begin with a set of strategies for task completion and these do not drastically change until the team reaches its midpoint transition. According to Gersick, this midpoint transition is the temporal period that marks the halfway point between the team's initial meeting and its official deadline. In terms of the measurement of team process within culturally diverse teams, this would suggest measuring around the midpoint transition, thereby taking advantage of natural breakpoints within the team when they will be the most receptive to feedback that flows from the assessment. While these recommendations are expected to hold within culturally diverse teams, different notions of time orientation and notions of time fluidity might cause these phases to be less clearly seen within culturally diverse teams.

Other insight into the timing of team performance measurement comes from stage models of team development (e.g., Morgan et al., 1986; Tuckman, 1965). For example, Tuckman proposed that teams progress through five stages of development: forming, storming, norming, performing, and adjourning. In terms of measurement with culturally diverse teams, the first four stages are the most relevant and will be briefly highlighted. Within the forming stage, team members begin to establish ground rules and members begin to get to know one another, but operation is primarily individual. This stage is often characterized by little agreement, varying degrees of commitment to the team, and unclear purpose. Within the storming phase members begin to communicate, but still primarily think of themselves as individuals as opposed to a part of a team. Conflict and power struggles often occur as the team begins to gain increased clarity and purpose. With respect to culturally diverse teams, surface-level diversity (e.g., social categories) is expected to primarily drive interaction during these first two stages and reinforce prior stereotypes. Therefore, measurement during this stage might focus on identifying the cultural stereotypes that are beginning to drive interaction such that existing stereotypes that are inaccurate can be corrected. The norming phase is characterized by a movement to the establishment of norms and the establishment of clear roles and responsibilities. Beginning with this phase, deep-level diversity (e.g., cultural differences in attitudes, values, beliefs, etc.) is expected to begin to more clearly drive member interaction. Finally, the performance phase is characterized by a clear goal and team vision where the focus is on collective task accomplishment. It is during this phase that the team processes that were argued to be important for culturally diverse teams would be expected to be the most developed and the impact of surface level cultural differences would be expected to have minimized. Process loss at this point is typically due to deep-level diversity. While some have criticized stage models due to their focus on linear as opposed to cyclical development, these models are still popular, and in terms of measurement, they highlight the notion that teams have different foci at different points in time. This is a factor that is important for the measurement of culturally diverse teams, because it guides what constructs might be best to measure at particular points in time.

12.3.6 Consideration 6: How to Make Sense of Responses (Indexing and Aggregation)?

The final component of a measure is the manner in which it is indexed and aggregated. The predominant practice within the team process literature is to collect data from individual team members using a team referent; however, as mentioned earlier there are small pockets of research that collect process data at the team level by asking the team, as a whole, to answer process questions. When data are collected from individuals using a team referent, issues of aggregation become paramount, while with the latter approach, aggregation is not an issue. When deciding on the appropriateness of aggregating individual responses to the team level and the method of indexing we must consider several benchmarks: (a) theory, (b) the manner in which the question is asked (e.g., does it refer to a team referent), and (c) empirical demonstration of greater within- than between-group agreement (see Kenny & LeVoie, 1985; Tesluk, Farr, & Klein, 1997 for statistics that allow this to be computed). Indexing refers to the statistic that is used to represent the construct (e.g., mean, variance, difference scores, minimum or maximum score). Traditionally the most common indexing method within the literature on team process has been to use a team average or mean. While the mean is still the most prominent indexing method, several recent authors have argued for the importance of really thinking about which index best represents the construct of interest (Kozlowski & Klein, 2000; Smith-Jentsch, 2009). This should be guided by theory as well as team and task characteristics. The degree of cultural diversity is one of the team characteristics that could impact indexing and aggregation.

Team researchers have argued that constructs emerge at the team level in one of two ways, composition or compilation (Kozlowski & Klein, 2000). Constructs that emerge via composition reflect situations in which the individual-level (i.e., lower level) variable is isomorphic with the team-level variable, such that the construct being assessed is essentially identical at both the individual and team levels. Given the restricted within-unit variance in this case, aggregation to the team level can be best represented by the sum or mean. Conversely, when constructs emerge via compilation, it is based on the idea that the configuration of different lower-level properties results in the higher-unit level property (i.e., team level construct; Kozlowski & Klein, 2000). Such constructs do not represent shared properties (i.e., they are not isomorphic) across levels, but instead are qualitatively different, such that the constructs are characterized by patterns. Given the variations in expectations within culturally diverse teams, it is likely that many team processes emerge via compilation and are not isomorphic across levels. This, in turn, implies that indexing is best represented not by the mean, but by variance, minimum or maximum, profile similarity, or other techniques that take into account this patterning (see Kozlowski & Klein, 2000)

12.4 Conclusion

Globalization is causing the physical boundaries between nations to be minimized and allowing organizations to bring in expertise regardless of physical location. The ability to bring in multiple experts to tackle the types of complex collaborative tasks that are indicative of today's workplace is alluring for organizations. This is one of many factors that have caused culturally diverse teams to be increasingly common within organizations. However, examples from several domains (i.e., sports, military, aviation, medicine) have shown that a team of experts is different than an expert team, and that the transformation from one to the other does not happen automatically. Adding cultural diversity to the mix, while offering potential advantages in terms of increased synergy, often makes the initial work that teams have to engage in more complex as members navigate hidden cultural landmines. Therefore, providing longitudinal assessments (and corresponding feedback) in terms of their ability to work together as a team is essential.

Within the current chapter we have begun to highlight a subset of the team processes that might be most likely to cause culturally diverse teams to derail and be ineffective. Thereby we have provided some initial insight into the types of team processes that measurement systems should be focusing on. We do not mean to say that these are the only important processes, as much work remains to be done at the intersection of the teams and cross-cultural literature bases, such that there is a much better understanding of how the various cultural orientations of the team members may interact—in some cases leading to synergy and in others fostering process loss. We have begun to highlight some of the important processes to measure, and we have also extracted what is known about building quality team performance measures and cross-cultural measurement to highlight a set of six considerations (and corresponding guidance where available) that those charged with developing measures for culturally diverse teams should be cognizant of in order to maximize the efficacy of developed measures. We hope that this provides not only food for thought, but the impetus for more research in this area.

Acknowledgements This work was supported in part by a grant from the National Aeronautics and Space Administration (#NNJ13HF08P). The views expressed in this work are those of the authors and do not necessarily reflect the organizations with which they are affiliated or their sponsoring institutions or agencies.

References

Adler, N. J. (1997). *International dimensions of organizational behavior* (3rd ed.). Cincinnati, OH: International Thomson Publishing.

Arman, G., & Adair, C. K. (2012). Cross-cultural differences in perception of time: Implications for multinational teams. *European Journal of Work and Organizational Psychology, 21*(5), 657–680.

Arthur, W., Jr., Villado, A. J., & Bennett, W., Jr. (2012). Innovations in team task analysis: Identifying team-base task elements, tasks, and jobs. In M. A. Wilson, W. Bennett, S. G. Gibson, & G. M. Alliger (Eds.), *Handbook of work analysis: Methods, systems, applications and science of work measurement in organizations* (pp. 641–661). New York, NY: Routledge/Taylor and Francis Group.

Bass, B. M. (1997). Does the transactional–transformational leadership paradigm transcend organizational and national boundaries? *American Psychologist, 52*(2), 130–139.

Berger, M. (Ed.). (1996). *Cross cultural team building: Guidelines for more effective communication and negotiation*. McGraw-Hill Book Company Limited.

Boyd, J. E., Kanas, N. A., Salnitskiy, V. P., Gushin, V. I., Saylor, S. A., Weiss, D. S., et al. (2009). Cultural differences in crewmembers and mission control personnel during two space station programs. *Aviation, Space and Environmental Medicine, 80*(6), 532–540.

Brannick, M. T., Levine, E. L., & Morgeson, F. P. (2007). *Job and work analysis: Methods, research, and applications for human resource management*. Thousand Oaks, CA: Sage Publications.

Brodbeck, F. C., Frese, M., Akerblom, S., Audia, G., Bakacsi, G., Bendova, H., et al. (2000). Cultural variation of leadership prototypes across 22 European countries. *Journal of Occupational and Organizational Psychology, 73*(1), 1–29.

Brown, A. L., Adams, B. D., Famewo, J. J., & Karthaus, C. L. (2008). *Trust in culturally diverse teams* (Contractor Report No. DRDC Toronto CR 2008-097). Toronto, Canada: Defence Research and Development.

Burke, C. S. (2005). Team task analysis. In N. Stanton, A. Hedge, K. Brookhuis, E. Salas, & H. Hendrick (Eds.), *Handbook of human factors and ergonomics methods* (pp. 526–535). Boca Raton, FL: CRC Press.

Carson, J. B., Tesluk, P. E., & Marrone, J. A. (2007). Shared leadership in teams: An investigation of antecedent conditions and performance. *Academy of Management Journal, 50,* 1217–1234.

Cherrie, S. F. (1997). Multinational mine-strike recovery operation. *Engineer, 27*(3), 6.

Choi, I., & Nisbett, R. E. (2000). Cultural psychology of surprise: Holistic theories and recognition of contradiction. *Journal of Personality and Social Psychology, 79*(6), 890–905.

Conyne, R. K., Wilson, F. R., Tang, M., & Shi, K. (1999). Cultural similarities and differences and differences in group work: Pilot study of a U.S.-Chinese task group comparison. *Group Dynamics: Theory, Research and Practice, 3*(1), 40–50.

Cooke, N., Salas, E., Cannon-Bowers, J. A., & Stout, R. (2000). Measuring team knowledge. *Human Factors, 42,* 151–173.

David, E. M., Rubino, C., Keeton, K. E., Miller, C. A., & Patterson, H. N. (2011). *An examination of cross-cultural interactions aboard the International Space Station* (NASA Technical Report No. TM-2011-217351). Human Research Program, Behavioral Health and Human Performance Element, Space Medicine Division.

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: A meta-analysis. *Journal of Applied Psychology, 95*(1), 32–53.

Gelfand, M. J., Raver, J. L., & Ehrhart, K. H. (2002). Methodological issues in cross-cultural organizational research. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial/organizational psychology* (pp. 216–246). Malden, MA: Blackwell Publishing.

Gersick, C. J. G. (1988). Time and transition in work teams: Toward a new model of group development. *Academy of Management Journal, 31*(1), 9–41.

Gibson, C. B., & Zellmer-Bruhn, M. (2001). Metaphors and meaning: An intercultural analysis of the concept of teamwork. *Administrative Science Quarterly, 46*(2), 274–303.

Gibson, C. B., & Zellmer-Bruhn, M. (2002). Minding your metaphors: Applying the concept of teamwork metaphors to the management of teams in multicultural contexts. *Organizational Dynamics, 31*(2), 101–116.

Gomez, C., Kirkman, B. L., & Shapiro, D. L. (2000). The impact of collectivism and in-group/out-group membership on the evaluation generosity of team members. *Academy of Management Journal, 43*(6), 1097–1106.

Gudykunst, W. B., Matsumoto, Y., Ting-Toomey, S. T. E. L. L. A., Nishida, T., Kim, K., & Heyman, S. (1996). The influence of cultural individualism-collectivism, self construals, and individual values on communication styles across cultures. *Human Communication Research, 22*(4), 510–543.

Hall, E. T., & Hall, M. R. (1990). *Understanding cultural differences*. Boston, MA: Intercultural Press.

Helmreich, R. (2000). Culture and error in space: Implications from analog environments. *Aviation, Space and Environmental Medicine, 71*(9–11), 133–139.

Helmreich, R., & Merritt, A. (1998). *Culture at work in aviation and medicine: National, organizational and professional influences*. Brookfield, VT: Ashgate.

Hofstede, G. (1980). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Beverly Hills, CA: Sage Publications.

Hofstede, G. (1993). Cultural constraints in management theories. *Academy of Management Perspectives, 7*(1), 81–94.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (Eds.). (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Thousand Oaks, CA: Sage.

House, R. J., Hanges, P. J., Ruiz-Quintanilla, S. A., Dorfman, P. W., Javidan, M., Dickson, M., et al. (1999). Cultural influences on leadership and organizations: Project GLOBE. In W. H. Mobley, M. J. Gessner, & V. Arnold (Eds.), *Advances in global leadership* (pp. 171–233). Stamford, CT: JAI Press.

Humes, M., & Reilly, A. H. (2007). Managing intercultural teams: The eOrganization exercise. *Journal of Management Education, 32*(1), 118–137.

Jung, D. I., & Yammarino, F. J. (2001). Perceptions of transformational leadership among Asian Americans and Caucasian Americans: A level of analysis perspective. *The Journal of Leadership Studies, 8*(1), 3–21.

Kanas, N., Sandal, G., Boyd, J. E., Gushin, V. I., Manzey, D., North, R., … Wang, J. (2009). Psychology and culture during long-duration space missions. *Acta Astronautica, 64*(7–8), 659–677.

Kealey, D. J. (2004). Research on intercultural effectiveness and its relevance to multicultural crews in space. *Aviation, Space, and Environmental Medicine, 75*(7, Supplement 1), C58–C64.

Kenny, D. A., & LaVoie, L. (1985). Separating individual and group effects. *Journal of Personality and Social Psychology, 48,* 339–348.

Klein, H. A. (2004). Cognition in natural settings: The cultural lens model. In M. Kaplan (Ed.), *Advances in human performance and cognitive engineering research* (Vol. 4, pp. 249–280). Bingley, UK: Emerald Group Publishing.

Klein, H. A., Klein, G., & Mumaw, R. J. (2001). *Culturally sensitive aviation demands* (Technical Report Prepared for the Boeing Company under General Consultant Agreement 6-111-10A-0112). Fairborn, OH: Wright State University.

Kluckhohn, F. R., & Strodbeck, F. L. (1961). *Variations in value orientations*. Evanston: Row, Peterson and Company.

Kirkman, B. L., & Shapiro, D. L. (2001). The impact of team members' cultural values on productivity, cooperation, and empowerment in self-managing work teams. *Journal of Cross-Cultural Psychology, 32*(5), 597–617.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent properties. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco: Jossey-Bass.

Lane, H., & DiStefano, J. (1992). *International management behavior: From policy to practice*. Boston, MA: PWS-Kent.

Langan-Fox, J., Anglim, J., & Wilson, J. R. (2004). Mental models, team mental models, and performance: Process, development, and future directions. *Human Factors and Ergonomics in Manufacturing & Service Industries, 14*(4), 331–352.

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. (2008). A meta-analysis of teamwork processes: Tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology, 61,* 273–307.

Li, H. Z. (1999). Communicating information in conversations: A cross-cultural comparison. *International Journal of Intercultural Relations, 23*(3), 387–409.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team process. *Academy of Management Review, 26*(3), 356–376.

Mathieu, J. M., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team effectiveness: 1997–2007: A review of recent advancements and a glimpse into the future. *Journal of Management, 34*(3), 410–476.

Mohsinsky, D. (2000). Acculturation gap and granparent's perceptions of their grandchildren in families of refugees from teh former Soviet Union. Unpublished undergraduate honor's thesis. University of Maryland, College Park.

Morgan, B. B., Jr., Glickman, A. S., Woodard, E. A., Blaiwes, A. S., & Salas, E. (1986). *Measurement of team behaviors in a Navy environment* (Tech Report No. NTSC TR-86-014). Orlando, FL: Naval Training Systems Center.

Newman, K. L., & Nollen, S. D. (1996). Culture and congruence: The fit between management practices and national culture. *Journal of International Business Studies, 27,* 753–779.

Oertig, M., & Buergi, T. (2006). The challenges of managing cross-cultural virtual project teams. *Team Performance Management: An International Journal, 12*(1/2), 23–30.

Pillai, R., & Meindl, J. R. (1998). Context and charisma: A "meso" level examination of the relationship of organic structure, collectivism, and crisis to charismatic leadership. *Journal of Management, 24*(5), 643–671.

Rentsch, J. R., & Woehr, D. J. (2004). Quantifying congruence in cognition: Social relations modeling and team member schema similarity. In E. Salas & S. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 11–31). Washington, DC: American Psychological Association.

Rosen, M. A., Schiebel, N., Salas, E., Wu, T. S., Silvestri, S., & King, H. B. (2012). How can team performance be measured, assessed, and diagnosed. In E. Salas & K. Frush (Eds.), *Improving patient safety through teamwork and team training* (pp. 59–79). Oxford, UK: Oxford University Press.

Salas, E., Burke, C. S., Wilson-Donnelly, K. A., & Fowlkes, J. E. (2004). Promoting effective leadership within multicultural teams: An event-based approach. In D. Day, S. J. Zaccaro, & S. M. Halpin (Eds.), *Leader development for transforming organizations* (pp. 293–323). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Salas, E., Shuffler, M. L., Thayer, A. L., Bedwell, W. L., & Lazzara, E. H. (2015). Understanding and improving teamwork in organizations: A scientifically based practical guide. *Human Resource Management, 54*(4), 599–622.

Sato, E. (2016). *Non-cognitive factors, culture, and fair and valid assessment of culturally- and linguistically-diverse learners*. Presentation given to the National Council on Measurement in Education.

Schweder, R., & Bourne, E. J. (1982). Does the concept of the person vary cross-culturally? In A. J. Marsella & G. M. White (Eds.), *Cultural conceptions of mental health and therapy* (pp. 97–137). Dordrecht: Reidel.

Sims, D. E., & Salas, E. (2007). When teams fail in organizations: What creates teamwork breakdowns? In J. Langan-Fox, C. L. Cooper, & R. J. Klimoski (Eds.), *Research companion to the dysfunctional workplace: Management challenges and symptoms* (pp. 302–318). Cheltenham, UK: Edward Elgar Publishing Limited.

Smith-Jentsch, K. A. (2009). The devil is in the details. In E. Salas, G. F. Goodwin, & C. S. Burke (Eds.), *Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches* (pp. 491–508). New York, NY: Routledge/Taylor & Francis Group.

Tesluk, P. E., Farr, J. L., & Klein, S. R. (1997). Influences of organizational culture and climate on individual creativity. *The Journal of Creative Behavior, 31*(1), 27–41.

Ting-Toomey, S. (1999). *Communicating across cultures*. New York, NY: Guildford Press.

Triandis, H. C. (1993). Collectivism and individualism as cultural syndromes. *Cross-Cultural Research, 27*(3/4), 155–180.

Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological Bulletin, 63*(6), 384–399.

Watson, W. E., Johnson, L., & Zgourides, G. D. (2002). The influence of ethnic diversity on leadership, group processes, and performance: An examination of learning teams. *International Journal of Intercultural Relations, 26,* 1–16.

Chapter 13
Inclusive Design of Collaborative Problem-Solving Tasks

Markku T. Hakkinen and Jason J.G. White

Abstract The design of collaboration tasks to be inclusive of people with disabilities raises unique practical challenges and opportunities for research. In this chapter, we review the context established by regulations and international standards in which efforts to develop accessible collaboration software are situated. Prior work in the design of such systems is briefly surveyed, and we identify unsolved problems that remain in a field which is yet to become the subject of sustained research and implementation experience.

Keywords Assistive technology · Accessibility · Regulation · International standards · Inclusiveness · Collaborative problem solving

13.1 Introduction

The key role that information technology plays in supporting collaborative problem-solving activities presents both opportunities and challenges for the inclusion of individuals with disabilities. Collaborative problem solving is by nature a human activity, and as such, technology functions as a mediator between humans engaged in a common task. Abrami and Bures (1996) and Schneiderman, Alavi, Norman, and Borkowski (1995) recognized the importance of considering the needs of individuals with disabilities in collaborative learning, but the results in the interim have been mixed. While technologies can enable inclusion, they can also create barriers when a given technology effectively excludes a specific population through a mismatch between the technical capabilities of the system and the sensory, cognitive, or physical requirements of the user of that system (Lazar & Jaeger, 2011). Legislation has emerged in many countries that codifies inclusion of

M.T. Hakkinen (✉) · J.J.G. White
Educational Testing Service, Princeton, NJ, USA
e-mail: mhakkinen@ets.org

J.J.G. White
e-mail: jjwhite@ets.org

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_13

individuals with disabilities and places specific technical requirements on information and communications technologies to support, for example, assistive technologies, such as screen-reading software used by those with visual impairments. Creating systems that are usable by a diverse community of end users is achieved through applying accessible and universal design practices early in the design and implementation cycles. The concept that new systems can be "born accessible" (Wentz, Jaeger, & Lazar, 2011) is an emerging paradigm, but one that faces many challenges, among them a lack of the necessary knowledge and skills among technologists and developers. Failing to address accessibility at the outset can result in costly remediation and reengineering to bring a system into conformance with accessibility legislation and standards.

In this chapter we will first introduce the role of universal design and assistive technologies in making information and communications technologies accessible to people with disabilities, then give an overview of inclusive design for collaborative problem solving through an examination of the legislation and technical standards applicable to developing collaborative problem solving systems, provide a brief summary of research in accessible collaboration, and close with a series of challenges and opportunities for further research.

13.2 Understanding Disabilities and Assistive Technologies

Disability is not a small or marginal phenomenon. According to the World Health Organization, 15% of the global population has a disability (World Health Organization, 2015). This population can include persons with any of a broad range of physical, sensory, cognitive, psychiatric, and learning disabilities. With the current global population at 7.3 billion people (U.S. Census Bureau, 2016), this translates into approximately 1 billion people living with some form of disability, and in many cases, with more than one functional impairment or limitation. These disabilities can and do pose challenges for inclusion in everyday activities, including education and employment. With the ongoing transformation to a digital world, technology-based products and services have changed how we communicate and share information. However, for many with disabilities, the ability to interact with common, everyday technology can be a challenge, and has resulted in what has been termed a digital divide affecting people with disabilities (Waddell, 1999) or more appropriately, a "disability divide" (Dobransky & Hargittai, 2006; Solomon, 2000). A key challenge area is for individuals with sensory disabilities, such as hearing or visual impairments. Visual and auditory presentation modalities, inherent in computer-based communications and collaboration platforms, can raise specific barriers for people who may be blind or deaf. Further, emerging gestural and touch interfaces can pose challenges for people with physical or mobility barriers.

For individuals with disabilities, two approaches, often complementary, have emerged for overcoming barriers to access: universal design, and assistive technologies (Vanderheiden, 1998). Universal design defines an approach in which systems are designed at the outset directly to support a broad range of abilities and disabilities. While no system can truly provide universal access to all users, the design approach can nonetheless improve overall usability for a broad range of users. In contrast, assistive technologies have been traditionally viewed as software and hardware add-ons that provide accessibility to a system that itself would otherwise not be accessible to a person with a specific disability. Examples of assistive technologies include screen readers (translating visual applications into spoken and braille presentations for blind and visually impaired users), screen magnification (enlarging a visual application for users with low vision), switch access interfaces (enabling users who cannot use a mouse or keyboard to create inputs), and augmentative communication tools (allowing users with speech impairments to communicate).

A growing number of off-the-shelf products, such as smart phones, tablets, set-top boxes, and personal computers, now blend universal design with built-in assistive technologies, significantly increasing accessibility for many users with disabilities. A user with a visual impairment, for example, can simply turn on a device's accessibility features to enable screen reading or magnification capabilities. For systems without built-in features, many assistive applications can be downloaded (some for free or at low cost) and installed.

While the presence of universal design features and assistive technologies enables access for users with disabilities, the applications and content that a user interacts with must also support the use of assistive technologies. This support is achieved through conformance with accessibility standards and best practices, with applications and content designed to provide access through a combination of platform capabilities and software code that adheres to accessibility standards (Brunet, Feigenbaum, Harris, & Laws, 2005). The importance of accessibility standards is underlined by the incorporation of those standards in national and international legislation to ensure access for persons with disabilities.

13.3 Legislation and Guidelines

In the contemporary policy environment, interactive collaboration software is subject to a complex combination of regulations and international standards in regard to its accessibility to people with disabilities. Although a detailed treatment of applicable standards and regulatory requirements would exceed the scope of this chapter, an analysis can nevertheless be given that seeks to identify and classify germane sources of policy.

The relevant international technical standards can be regarded as falling into two general categories. First, there are specific standards that define the details of technologies that may be used in the implementation of collaboration software.

For example, in developing accessible Web applications, including collaborative tools, it is typically necessary to use the facilities defined in the Accessible Rich Internet Applications (WAI-ARIA) 1.0 specification (World Wide Web Consortium [W3C], 2014). This standard enables custom-built interactive user interface controls to be made accessible via screen readers and, potentially though not yet in practice, by other assistive technologies. It also provides for landmarks to be defined which allow a screen reader user to move the navigational focus to specific parts of a Web application, for example a search form or the main content area. These landmarks identify important locations that enable a screen reader user to gain an overview of the main elements of the user interface, to identify key portions of the content, and to jump directly to these points for purposes of efficiently reading and interacting with the application. Similarly, Timed Text Markup Language 1.0, second edition (W3C, 2013), and WebVTT (W3C, 2016; merely a draft community group report as of December 2015, but nonetheless already implemented in Web browsers), are technologies designed for the specific purpose of including text tracks in video content. In particular, either of these technologies can be used to provide captions for the auditory track of a video, thereby meeting the needs of people who are deaf or who have a hearing impairment.

Second, there are more abstract and general technical standards that establish the requirements which need to be satisfied in order for documents or applications to be accessible. The most widely cited of these standards is the Web Content Accessibility Guidelines (WCAG) 1.0 and more recently 2.0 (W3C, 2008), the scope of which is sufficiently broad to encompass highly interactive Web-based applications, including collaboration software that is implemented by means of Web technologies. WCAG 2.0 articulates four general principles of Web accessibility, which assert that accessible content, including applications, must be "perceivable," "operable," and "understandable" by people with disabilities as well as "robust" in the sense of supporting compatibility with a variety of Web browsers and assistive technologies. Under each of these four principles appear more specific guidelines, each of which is associated with success criteria–testable assertions that must be true in order for the requirement expressed in the accompanying guideline to be met. The success criteria are ranked according to three levels of conformance, in which each successive level establishes a higher degree of general accessibility. The second level of conformance, level AA, is notable for having been cited internationally in current and proposed regulations and government policies (Rogers, 2016). WCAG 2.0 and its predecessor, WCAG 1.0, have also been referred to in judicial and administrative proceedings, for example in settlement agreements reached by the United States Department of Justice in cases of alleged discrimination under the Americans with Disabilities Act (Department of Justice, 2014, 2015), and in Canadian federal court—see *Jodhan v. Canada* (2010).

The legal context relevant to the accessibility of collaboration software varies greatly between countries. It is also likely to evolve significantly in response to regulatory changes, judicial determinations, and shifts in administrative policy. Identifying which laws are applicable to collaboration software under particular circumstances requires a meticulous legal analysis. In general, the law regulates the accessibility of information and communications technology in two distinct ways. First, the entities responsible for the creation, distribution, or use of the technologies may be subject to a general prohibition of discrimination against people with disabilities. The precise conditions under which the prohibition applies depend on the details of the legislation in force in a particular jurisdiction. The establishment of such general prohibitions can be exemplified by the nondiscrimination provisions of the Americans with Disabilities Act 1990 (as amended) in the United States, the Equality Act 2010 in the United Kingdom and the Disability Discrimination Act 1992 (Commonwealth) in Australia. Second, the law may require specific standards of accessibility or nondiscrimination to be satisfied with respect to a defined class of organizations or a particular domain of activity. For example, in the United States, Section 508 of the Rehabilitation Act of 1973 and its accompanying regulations (Electronic and Information Technology Accessibility Standards, 2011) establish a technical standard of accessibility which is to be satisfied, with certain exceptions, by software developed or procured by the federal government. This accessibility requirement, and its counterparts elsewhere, give rise to an economic incentive by opening the public sector market only to vendors whose products and services meet a prescribed standard.

Collaboration software used in educational contexts may also be subject to specific regulations. An example can be found in the Disability Standard for Education 2005 (Commonwealth of Australia, 2005), a regulation established by the Australian government pursuant to the Disability Discrimination Act, 1992 that requires “reasonable adjustments” to be made to courses, programs, and curricula, including learning experiences and assessment, in order to enable a student with a disability to participate “on the same basis” as a student without a disability. Furthermore, collaboration software that involves real-time interaction among participants by way of voice, video, or text messages may in some circumstances be regulated as a telecommunication service. The U.S. Federal Communications Commission has issued regulations under Section 617 of the Communications Act of 1934 that require “advanced communications services” to be accessible to people who have disabilities affecting hearing, vision, speech, motor, and cognitive abilities.

As this discussion has demonstrated, the accessibility of collaboration software may be mandated by a variety of regulations, each possessing unique conditions of applicability and asserting different substantive requirements. This regulatory diversity is particularly apparent from consideration, as here, of an international sample of relevant laws. While conforming to international accessibility standards provides no guarantee that a given legal requirement is met unless the law itself designates the standards as sufficient, implementing such standards is an important and valuable measure to be taken in the design and development of collaborative problem-solving systems.

The fundamental principle that underpins the regulations and policies here described is that of enabling people with disabilities to use information and communications technologies on an equal footing to their counterparts who do not have disabilities. Conforming to international standards, in particular WCAG 2.0 at

Level AA, and observing established practices of accessible software development are important measures that authors of interactive collaboration applications can take so as to give effect to this principle. Amid the differences and uncertainties characteristic of the regulatory environment, it is essential that the unifying principle of equal treatment be borne constantly in mind, and that attention be paid to the concrete, practical details of how people with different capabilities, some of whom use a variety of assistive technologies, would use the software that is ultimately created. International standards should likewise be applied with a view to their purpose rather than in a legalistic or technocratic manner that seeks to meet the letter of the requirements while disregarding their objectives and the effects of decisions taken in the design and implementation of the software on users with disabilities. These remarks, which in no way constitute advice regarding compliance with legal requirements, are of particular significance in the development of collaboration software, owing to the limited research and practical experience in making these collaborative environments accessible that can serve to guide decisions made in the construction of new interactive collaboration systems.

13.4 Development of Accessibility Standards as Collaborative Problem Solving

The development of technical accessibility standards in recent decades has focused primarily, though by no means exclusively, on the World Wide Web. In the context of collaborative educational applications, moreover, Web technologies occupy a fundamental role. Stand-alone collaborative applications and collaborative components of larger Web-based tools can generally be implemented by means of standard protocols, content formats, and programmatic interfaces. Equally, Web technologies may be used in the development of software designed to be packaged and deployed on mobile devices. Hence the centrality of Web technologies and standards to relevant practices of application development justifies the emphasis that is apparent in the discussion that follows.

The World Wide Web Consortium (W3C) is the principal organization responsible for the creation, revision, and dissemination of Web-related technical standards generally and of accessibility standards specifically. Other organizations, for example the Instructional Management System (IMS) Global Learning Consortium in the field of educational technology, have also played a part, but the underlying standards that have shaped technical developments and public policy are those of the W3C. As formalized in W3C's Process Document (W3C, 2015), and as carried out in practice, the creation of technical standards is a cooperative activity through which participants strive to reach consensus among themselves and with a broader community of reviewers. Documents are developed by working groups comprising representatives of the Consortium's member organizations and invited technical experts, drafts are periodically published to solicit wider review, and implementation experience is collected—all in a context characterized by collaborative problem solving.

Since the establishment of the W3C's Web Accessibility Initiative in 1997, these processes of collaborative problem solving in the development of often complex technical standards have been inclusive of participants with disabilities. The inclusion of individuals with diverse access requirements has been facilitated by the use of standards-based collaboration tools—a combination of Internet mailing lists, teleconferences, Internet Relay Chat (IRC) for synchronous discussion, online surveys to collect comments, and revision control software to manage collaborative development of documents. To the extent that the W3C has succeeded in engaging participants with disabilities, it has done so by adhering to tools and practices which are broadly supportive of different operating systems, assistive technologies, and types of user interface. In order to be accessible to people with disabilities, collaborative learning applications should likewise be designed with flexibility in mind, while conforming to relevant technical standards.

13.5 Approaches to Accessibility in Collaborative Learning Environments

The challenge in designing inclusive collaborative problem-solving systems is to ensure that each participant is able to perceive and understand available information and construct responses that serve as input to the collaborative process. The most basic and accessible form of information presentation and input is text, and it has been a key component in both research and practice. Text-based information is easy to create, through keyboards, alternative input devices, or speech recognition, and can be transformed by assistive technologies into spoken form via speech synthesis or tactile form via braille displays. Furthermore, its visual rendering, including the size and style of font used and the contrast between foreground and background colors, can be varied in accordance with a user's requirements, in some cases by way of operating system or browser-based controls that do not require the intervention of assistive technologies. The use of text for collaborative writing tasks for students who are deaf has been described by Batson (1993), and text has emerged as an important modality in collaborative learning for those with hearing impairments. Sign language (such as American Sign Language, or ASL) is a key means of discourse for many who are deaf, and the interspersing of text and sign language may be beneficial, for example, in science learning tasks (Lang and Steely, 2003). While text-to-speech synthesis is an effective and accepted, means for translating written text into a spoken form for those with visual impairments, the transformation of text into sign language using "signing avatars" remains an area of significant research, though it is generally seen as not ready for practical application (Kipp, Nguyen, Heloir, & Matthes, 2011). The difficulties which this transformation raises are attributable in large part to the fact that it requires the automatic translation of one language into another, for example, of English into ASL. The problem is thus similar in complexity to that confronting any attempt to automate the translation of text between languages, as may be desirable for instance to assist collaborators in an intercultural setting whose knowledge of a shared language is inadequate to the task which is to be performed. In the absence of substantial innovations in computational linguistics, however, direct textual communication (unmediated by automatic translation) remains the most broadly accessible communication medium for use in collaborative environments.[1]

In practice, collaborative problem-solving tasks involve more than just text-based interaction. Spatial information and tasks can be key to demonstrating a problem or its solutions. Nonverbal cues, such as facial expressions, gestures, or gaze may convey significant, real-time information that is not easily transformed by assistive technologies into a textual form. Accessibility researchers have been exploring different approaches to transforming spatial and nonverbal information into modalities suitable for those with visual impairments. Sonification is one such approach for conveying data and spatial information (Hermann, Hunt, & Neuhoff, 2011). Tanveer, Anam, Rahman, Ghosh, and Yeasin (2012), for example, proposed a sensory substitution system that dynamically transforms facial expressions into a sonified audio representation. Haptics, an approach that utilizes the sense of touch to convey information, is also emerging as a promising modality. Winberg (2006) described a system for cross-modal collaboration incorporating both auditory and haptic feedback, enabling a sighted and a visually impaired student to work together in performing sorting and handover tasks. The combination of multiple modalities, such as audio and haptics, with human interpreters to describe information not otherwise transformable, is being explored by Pölzer and Miesenberger (2014) for presenting nonverbal communication that occurs in collaborative brainstorming.

The conveying of nonverbal cues to participants who cannot perceive them directly in a communication remains a substantial challenge to the design of accessible collaboration systems. A further challenge results from the synchronous nature of the communication induced by collaboration tools, whereby long delays in responding to an interlocutor's contribution that may be necessitated by the use of assistive technologies (for example, alternative input devices) has the potential to have adverse effects upon the completion of collaborative tasks that are time-critical. On the positive side, the availability of asynchronous communication methods as a complement to real-time interaction can facilitate communication between people with different capabilities as well as participants from diverse linguistic backgrounds. This observation is well illustrated by the W3C's collaboration process, in which electronic mailing lists remain a central means of collaboration that complement, but have not been supplanted by synchronous means of interaction, namely teleconferences and IRC. Unusually slow responses resulting

[1] The accessibility advantage of textual communication stands as an independent reason for preferring it from the ground advanced by Hao, Liu, Von Davier, & Kyllonen (this volume), namely its effectiveness in reducing confounding factors that may otherwise interfere with the measurement of collaboration skills.

from the use of assistive technologies may also serve as indicators that a participant probably has a disability, thus raising concerns about the maintenance of privacy in a software-mediated collaborative environment.

Additional challenges for designers of collaborative problem solving systems arise from the need for participants to focus attention on the task to be performed, while being appropriately alerted to the communicative acts of collaborators. Nonintrusive indications of communication from fellow participants need to be provided in a sensory modality that is appropriate to the individual participant. It may also be advantageous to employ a separate sensory modality for communication from that used to interact with the entities manipulated in solving the problem at hand. For example, a user who is blind could interact with the problem-solving task via a braille or haptic display, while communicating auditorily with collaborators. It is an open research question whether and to what extent such an arrangement, where made feasible by the capabilities of the user, could improve task performance by reducing the participant's overall cognitive load. Moreover, the requirement that a problem-solving task be made simultaneously accessible to all collaborators imposes interesting requirements on the user interface. For instance, whereas drawing objects graphically in a workspace is easy and convenient for visual users, the semantics of what is drawn need to be captured explicitly by the system, for example in the form of metadata or descriptive text, in order to be conveyed to collaborators who are working in nonvisual modalities. A similar need is apparent with respect to the participation of users who are deaf or hearing-impaired in tasks requiring the generation and editing of sound. Of course, the availability of a synchronous channel of communication among collaborators enables them to assist each other in addressing problems of accessibility, but it is clearly desirable for the software to be designed to reduce the need for participants to engage in additional communication to overcome access barriers by encouraging, if not requiring, information to be entered in a form that can automatically be presented in different sensory modalities and in accordance with a variety of users' needs.

The steps used by the participants to interact with one another and with the problem-solving task provide useful data in assessing individual and group performance. For example, in the chapter by Hao et al. (this volume), text chatting is utilized as the collaboration channel and also serves as the primary metric. As we have described, such an approach should support the accessibility needs of a broad range of collaborators with disabilities. However, questions remain as to the impact of both disability and assistive technology on collaborative interaction, and how any impact would be accommodated. For example, students with physical disabilities may require adaptation of both software and hardware to support text entry. Specialized keyboard interfaces, for example, on-screen keyboards, have been developed, and in some cases embedded directly into computer operating systems (e.g., Microsoft Corporation, 2016). While on-screen keyboards can facilitate text entry, the combination of disability and interface typically can result in effective typing rates in the range of nine to 12 words per minute (Anson et al., 2006), and potentially as low as two words per minute (Tumlin & Heller, 2004). Differences in text entry rates between task participants when assistive technologies are used may introduce delays that impact communication and task completion performance. How such delays may impact the interaction and engagement of the participants, and what steps might be taken to mitigate the effect of those delays is an area for research. And, overall, a key question for those designing collaborative problem solving tasks is to understand how process data can take into account the presence of assistive technologies in a way that makes their use transparent when measuring performance. To answer that question, close cooperation is required between researchers in disability, accessibility, assistive technologies, and collaborative problem solving.

13.6 Conclusion

Researchers and developers in the field of collaborative problem solving face a challenge of building technology-based systems that do not exclude participation by individuals with disabilities. Existing and emerging legislation requiring accessibility, especially in the domains of education and employment, and technical standards that provide guidance, are a key starting point, as is the inclusion of experts in disability, accessibility, and assistive technologies in research and development efforts. Further, participatory design and evaluation by individuals with disabilities is vital in ensuring that resulting systems demonstrate usable accessibility and support true inclusion. As has been indicated in the preceding discussion, substantial research challenges remain to be overcome: the design of adequately inclusive collaboration software necessitates the development of new strategies for supporting accessibility that extend beyond the technical requirements and implementation techniques offered by existing international standards. The associated desire to analyze the interactive behavior of participants in order to acquire insight into their knowledge and skills raises further questions that remain to be investigated.

References

Abrami, P. C., & Bures, E. M. (1996). Computer-supported collaborative learning and distance education. *American Journal of Distance Education, 10*(2), 37–42.

Anson, D., Moist, P., Przywara, M., Wells, H., Saylor, H., & Maxime, H. (2006). The effects of word completion and word prediction on typing rates using on-screen keyboards. *Assistive Technology, 18*(2), 146–154.

Batson, T. (1993). ENFI research. *Computers and Composition, 10*(3), 93–101.

Brunet, P., Feigenbaum, B. A., Harris, K., & Laws, C. (2005). Accessibility requirements for systems design to accommodate users with vision impairments. *IBM Systems Journal, 44*(3), 445–466.

Commonwealth of Australia. (2005). *Disability standards for education 2005*. Government of Australia federal register of legislative instruments no. F2005L00767. Canberra, Australia: Australian government department of education and training.

Communications Act. (1934). 47 U.S.C. § 151 *et seq.*

Department of Justice. (2014). Settlement agreement between the United States of America and ahold U.S.A., Inc. and Peapod, LLC. Under the Americans with disabilities act. DJ No. 202-63-169. Retrieved June 7 2016, from https://www.justice.gov/file/163956/download

Department of Justice. (2015). Settlement agreement between the United States of America and EdX Inc. under the Americans with disabilities act. DJ No. 202-36-255. Retrieved June 7, 2016, from https://www.justice.gov/sites/default/files/opa/press-releases/attachments/2015/04/02/edx_settlement_agreement.pdf

Dobransky, K., & Hargittai, E. (2006). The disability divide in internet access and use. *Information, Communication & Society, 9*(3), 313–334.

Electronic and Information Technology Accessibility Standards. (2011). 36 C.F.R. pt 1194.

Hermann, T., Hunt, A., & Neuhoff, G. (Eds.). (2011). *The sonification handbook*. Berlin, Germany: Logos Verlag.

Jodhan v. Canada (Attorney General). (2010). FC 1197.

Kipp, M., Nguyen, Q., Heloir, A., & Matthes, S. (2011, October). Assessing the deaf user perspective on sign language avatars. In ACM (Ed.), *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 107–114). New York, NY: ACM).

Lang, H. G., & Steely, D. (2003). Web-based science instruction for deaf students: What research says to the teacher. *Instructional Science, 31*(4–5), 277–298.

Lazar, J., & Jaeger, P. (2011). Reducing barriers to online access for people with disabilities. *Issues in Science and Technology, 27*(2), 69–82.

Microsoft Corporation. (2016). *Use the on-screen keyboard (OSK) to type*. Retrieved June 7, 2016, from http://windows.microsoft.com/en-us/windows-8/type-with-the-on-screen-keyboard

Pölzer, S., & Miesenberger, K. (2014). Presenting non-verbal communication to blind users in brainstorming sessions. In K. Miesenberger, D. Fels, D. Archambult, P. Penaz, & W. Zagler (Eds.), *Computers helping people with special needs* (pp. 220–225). Cham, Switzerland: Springer International Publishing).

Rehabilitation Act, 29 U.S.C. § 794d (1973).

Rogers, M. (2016, March). *Government accessibility standards and WCAG 2*. Retrieved June 7, 2016, from http://www.powermapper.com/blog/government-accessibility-standards/

Shneiderman, B., Alavi, M., Norman, K., & Borkowski, E. Y. (1995). Windows of opportunity in electronic classrooms. *Communications of the ACM, 38*(11), 19–24.

Solomon, K. (2000). Disability divide. *The industry standard*, *3*.

Tanveer, M. I., Anam, A. S. M., Rahman, A. K. M., Ghosh, S., & Yeasin, M. (2012, October). FEPS: A sensory substitution system for the blind to perceive facial expressions. In ACM (Ed.), *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 207–208). New York, NY: ACM).

Tumlin, J., & Heller, K. W. (2004). Using word prediction software to increase typing fluency with students with physical disabilities. *Journal of Special Education Technology, 19*(3), 5–14.

U.S. Census Bureau. (2016). *U.S. and world population clock*. Retrieved June 7, 2016, from http://www.census.gov/popclock/

Vanderheiden, G. C. (1998). Universal design and assistive technology in communication and information technologies: Alternatives or complements? *Assistive Technology, 10*(1), 29–36.

Waddell, C. (1999, May). The growing digital divide in access for people with disabilities: Overcoming barriers to participation in the digital economy. White paper presented at the Understanding the Digital Economy: Data, Tools and Research conference, U.S. Department of Commerce, Washington, DC. (May 25–26, 1999). Retrieved June 7, 2016, from http://www.icdri.org/legal/the_growing_digital_divide.htm

Wentz, B., Jaeger, P. T., & Lazar, J. (2011). Retrofitting accessibility: The legal inequality of after-the-fact online access for persons with disabilities in the United States. *First Monday*, *16*(11). Retrieved June 7, 2016, from http://journals.uic.edu/ojs/index.php/fm/article/view/3666

Winberg, F. (2006). Supporting cross-modal collaboration: Adding a social dimension to accessibility. In D. McGookin & S. Brewster (Eds.), *Haptic and audio interaction design* (pp. 102–110). Berlin, Germany: Springer.

World Health Organization (2015). *Disability and health* World health organization fact sheet No. N°352. Retrieved June 7, 2016, from http://www.who.int/mediacentre/factsheets/fs352/en/

World Wide Web Consortium. (2008, December 11). *Web content accessibility guidelines (WCAG) 2.0.* Retrieved July 6, 2016, from http://www.w3.org/TR/WCAG20/

World Wide Web Consortium. (2013, September 24). *Timed text markup language 1 (TTML1)* (2nd ed.). Retrieved July 6, 2016, from http://www.w3.org/TR/ttaf1-dfxp/

World Wide Web Consortium. (2014, March 20). *Accessible rich internet applications (WAI-ARIA) 1.0.* Retrieved July 6, 2016, from http://www.w3.org/TR/wai-aria/

World Wide Web Consortium (2015, September 1). *World wide web consortium process document*. Retrieved June 7, 2016, from http://www.w3.org/Consortium/Process/

World Wide Web Consortium. (2016, July 4). *WebVTT: The web video text tracks format.* Retrieved July 6, 2016, from https://w3c.github.io/webvtt/

Part II
Modeling and Analysis

Chapter 14
Understanding and Assessing Collaborative Processes Through Relational Events

Aaron Schecter and Noshir Contractor

Abstract Effective teams are characterized by how skillfully they collaborate, coordinate, and interact while working towards their collective goals. These processes are inherently dynamic, and are best represented as a series of events (i.e. interactions). Whereas other methods for studying teams focus on the properties or structure of the group, an event-focused framework has potential to yield unique insights about the nature of collaboration. We therefore introduce the relational event framework, which is a statistical tool designed specifically to take advantage of event data. This method makes statistical inferences about what sequential patterns of collaboration ties form and how these patterns perform. In this chapter we introduce the reader to relational event modeling, including an overview of the necessary data, measures, and statistical models. We also provide insights on how this statistical technique can be utilized to assess and understand collaboration.

Keywords Relational events · Teams, team process · Social network analysis · Event history models · Structural signatures · Generative mechanisms

14.1 Introduction

Complex tasks are achieved through the efforts of highly productive, highly skilled teams. These specialized groups collaborate to produce outcomes well beyond the capabilities of any individual. Teams are present in all facets of life, from science to medicine, engineering to business. Increasingly, the sole practitioner cannot compete with a well-balanced, skillful group. However, we are often at a loss for explaining what makes a successful collaboration. The teams that work towards these collective tasks are living, breathing units with a character all their own, and

A. Schecter (✉) · N. Contractor
Northwestern University, Evanston, IL, USA
e-mail: aaronschecter2016@u.northwestern.edu

N. Contractor
e-mail: nosh@northwestern.edu

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_14

consequently studying them requires a level of sophistication on par with the complex nature of group behavior.

The implications of research on teams are straightforward; if we have a better team, we can expect better collaborations. Yet, this seemingly benign problem has no easy answer. Therefore we pose the simple question: why do some teams fail while others succeed? Often, it is not the inputs to the team that are problematic; by design, each individual can be highly skilled and/or knowledgeable of the task at hand. Rather, failure is rooted in poor interactions or a lack of "chemistry." While interpersonal chemistry has its own colloquial meanings, team chemistry is poorly understood. To truly understand what makes an effective team, we need to look deeper than inputs and outputs; specifically, it is the actions and interactions that unfold over time that represent the nature of a team.

A number of theories explain the nature and quality of team interaction and the relationship between collaborative skill and the final product. Kozlowski and Klein (2000) analyzed a team by its emergent properties, which are characteristics of the team and the individuals within it, as well as the configuration of attributes within the unit. For example, a team may be assessed by how much planning behavior they took part in during their collaboration. As an extension of this framework, Marks, Mathieu, and Zaccaro (2001) incorporated time into the analysis of teamwork. A collaboration will naturally move through phases, during which different types of interactions are necessary. For instance, at the beginning of a project, individuals may focus on defining goals and delegating roles, while during later phases they may focus more on coordinating specific tasks or managing the team's mood. More recently, Crawford and LePine (2013) proposed a configural view of teamwork suggesting that the pattern and structure of teamwork influences outcomes. For example, teams that centralize work around one individual may perform differently than teams that use a distributed collaboration.

Building on these frameworks for assessment of teamwork, Leenders, Contractor, and DeChurch (2015) have proposed a new paradigm to studying team process—relational events—that focuses on individual interactions over time. This approach frames collaboration and communication as a sequence of events; the unfolding of these events may be explained endogenously (prior actions taken by members of the team) or exogenously (changes in the team's environment). The relational event framework identifies emergent patterns of behaviors between individuals, as well as other factors which contribute to the generation of future actions (Butts, 2008). As a result, relational event models (REMs) answer the "what events should happen when" question posed by Marks et al. (2001), while also answering the "who talks to whom" question posed by Crawford and Lepine (2013). In contrast to prior approaches, a REM is multilevel, capturing in a single model the influences of individual, dyadic, triadic, and group-level characteristic on the dynamic unfolding of collaboration processes As a result, the assumption of homogeneity, both among team members and over time, is no longer needed.

In this chapter, we describe the relational event framework and illustrate how it can be applied to the assessment of collaboration. In particular, we specify the data structure required for this type of analysis and describe the development of event-based statistics for testing hypotheses. Next, we give a brief overview of how to fit relational event models and how to use these results to assess a collaborative effort. Finally, we give a brief example of a scenario in which REM is applied.

14.2 The Relational Event Framework

14.2.1 What Are Relational Events?

A relational event is any interaction or behavior that originates from an individual towards another individual or object (Butts, 2008). Relational events are encoded as units of data that include relevant information such as the sender, target, and time of the event. Additional information such as the type of event (e.g., phone call or text message), weight (Foucault Welles, Vashevko, Bennett, & Contractor, 2014), or valence (e.g., positive or negative interaction; Brandes, Lerner, & Snijders, 2009) may be observed and recorded (Marcum & Butts, 2015). A full relational event dataset is effectively a transcript of exactly what transpired during the course of collaboration.

Relational events may be applied in a number of different contexts. Perhaps the simplest example of such a behavioral event is a message, sent from one individual to another. For an example of a series of relational events in a three-person project group, see Table 14.1.

Table 14.1 could be converted to an event sequence in a straightforward fashion: $e_1 = (a, b, t_1), e_2 = (b, a, t_2),\ e_3 = (c, \{a, b\}, t_3)$. This process can be repeated for the whole dataset. However, events are not confined to messages. For example, events may be directed from an individual to a task or tool. Quintane, Conaldi, Tonellato, and Lomi (2014) used relational events to model the interactions between software developers and blocks of code over time. Vu, Pattison, and Robins (2015) studied the clicking behavior of students using online course material, as well as their interaction with chat rooms. Alternatively, events may be egocentric (i.e., focused on one individual); Marcum and Butts (2015) used this version of the model to track the behaviors of elderly individuals throughout the course of a day.

Table 14.1 Sample relational event sequence

Time (PM)	Sender	Receiver	Message
2:01:00	Adam	Bob	Did you finish your section yet?
2:01:05	Bob	Adam	No, not yet
2:01:14	Christina	Adam, Bob	I finished mine, can I help either of you?

14.2.2 How Are Relational Events Applied?

Relational event sequences differ from other social network techniques such as exponential random graph models (ERGMs; Lusher, Koskinen, & Robins, 2012) or stochastic actor-oriented models (SAOMs; Snijders, 1996). In ERGMs, the structure of a single graph is analyzed. The structure of ties between individuals is determined to be more or less prevalent than we would expect in a random graph. ERGMs are useful for studying structure of network ties that are relatively enduring states (such as trust) captured by concepts such as centralization or multiplexity (simultaneous occurrence of multiple ties), but are not suited for studying ties that are episodic events (such as a chat message). Snijders and colleagues modeled the evolution of network dynamics via a Markov process, with the state transitions dependent on the current network. These so-called SOAMs introduce time into the analysis of social networks. The models are actor-oriented because actors—who choose to create, maintain, and dissolve ties based on their current position within the network—drive changes within the network. These models are particularly appropriate when a snapshot of the network data is collected at discrete time intervals (such as a day, month, or year), but the underlying process cannot be observed.

Relational event models expand on both of these modeling frameworks to accommodate interaction data that is completely observable, and increasingly available, such as online chat logs or transcripts of conversations. Relational event data are used to posit what Leenders et al. (2015) termed as a sequential structural signature (SSS), which is a dynamic analog to the statistics used in ERGMs. SSSs are sequences of relational events that unfold in a particular pattern and are designed to represent theoretically interesting behavior sequences. SSSs characterize interactions of various types at multiple levels. In particular, they may be at the ego level, the dyad level, the triad level, or beyond. Additionally, SSSs can incorporate attributes of the actors, as well as the relations themselves.

To illustrate the notion of an SSS, we present a simple example. Preferential attachment is the tendency for individuals to communicate with others who have previously been epicenters of interaction (Barabási & Albert, 1999). Put simply, as individual A increasingly sends and receives messages from individual B, then individual C becomes increasingly likely to send a message to B. This mechanism captures the extent to which popularity drives future communication. In Fig. 14.1, we illustrate the preferential attachment SSS; solid lines represent past communication, while dashed lines represent the potential new communication. Arrows indicate directionality.

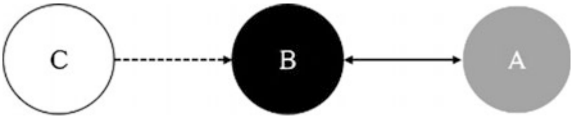


Fig. 14.1 Visual representation of preferential attachment SSSs

We now explain how to mathematically operationalize a signature such as the one presented in Fig. 14.1. Let n_{ijt} be the number of messages sent from i to j up to time t. As we stated in our description of preferential attachment, this signature represents an individual's level of activity, relative to the rest of the network. We provide a formula below (assuming N individuals):

$$s_{PA}\left(C,B,t\right)=\frac{\sum_{k=1,\ldots,N}n_{Bkt}+\sum_{k=1,\ldots,N}n_{kBt}}{\sum_{l=1,\ldots,N}\left(\sum_{k=1,\ldots,N}n_{lkt}+\sum_{k=1,\ldots,N}n_{ljt}\right)}.$$

The measure $s_{PA}\left(C,B,t\right)$ is the specific value of preferential attachment between sender C and receiver B at time t. The numerator is a sum of all incoming and outgoing messages involving node B up to the present time. The denominator is the sum of all messages sent and received in the network between any pair (l,k).

While the structure presented is straightforward, significantly more complex signatures can be developed. For instance, consider the case of two individuals collaborating on a software project. Let A and B be the individuals, and X is the software project they are considering working on. We represent this situation in Fig. 14.2. The solid line indicates that B has previously worked on the project, and that A and B have been communicating. The dashed line represents A's propensity to subsequently engage with the software project to potentially "redo" something just done by B.

We let the shading of A and B in Fig. 14.2 represent their relative experience; the grey circle represents the more knowledgeable member of the team. We would like to operationalize a statistic that captures the propensity for A to work on something B has already worked on, based on their prior communication, B's prior activity, and their relative skill difference. Using the same n_{ijt} notation as before and letting z_A denote the skill of individual A, we may create the following measure:

$$s_W\left(A,X,t\right)=n_{BXt}\times\left(n_{ABt}+n_{BAt}\right)\times\left(z_A-z_B\right).$$

This statistic will be large and positive if B has worked on software X more frequently, A and B have frequently communicated, and A is more skilled. If A becomes less likely to work on the software as s_W increases, then we would say that

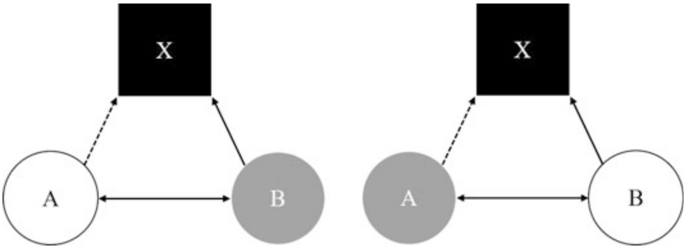


Fig. 14.2 Visual representation of communication and action

A has confidence in B’s abilities to get the work done. Alternatively, if A becomes more likely to engage X, then we might infer that A lacks confidence in B’s work, and decides to revise the item.

This approach to generating SSSs and operationalizing them can be applied to virtually any setting in which trace data are available. As with ERGMs or SAOMs, a visualization of the desired structure can be created, and accumulated interactions are used to represent the intensity of the hypothesized links. Attributes of the relationships or of the nodes themselves are easily incorporated, as illustrated above. The choice of statistics computed is based on theoretically motivated explanations for the emergence of events. Current research on relational events has used extensions of common signatures from ERGM or SAOM. Butts (2008) and Brandes et al. (2009) also provided a template for generating statistics. In general, the number and complexity of the terms are largely dependent on the theoretical explanation posited, as well as the context and availability of the data.

14.2.3 How Do We Fit Relational Event Models?

The foundation of REM is the specification of the rate function. The rate of an event represents its pace over time; more frequent events have a higher likelihood of occurring, relative to events with a lower rate. Event history analysis applies survival modeling to event data, and represents the event rate with a hazard function (see, for example, Blossfeld & Rohwer, 1995). The hazard rate for an event is the instantaneous likelihood of the action occurring, given its previous nonoccurrence. To account for the time between events, the survival function is used. The survival function is the likelihood that an event does not occur during a particular timespan. Survival functions may be directly computed from the hazard rate. As a result, determining a functional form for the hazard rate allows us to explicitly model a relational event sequence.

Butts (2008) defined the hazard rate λ for a relational event to be an exponential function of a linear combination of sufficient statistics s and rate parameters θ. The sufficient statistics are simply mathematical representations of SSSs, as discussed previously. The rate parameters are analogous to the parameters of a logistic regression model; the sign and significance indicates what effect the corresponding pattern has on future events. The functional form of the hazard rate is as follows:

$$\lambda_{ij}(t;\,\theta) = \exp\left(\sum_{p=1,\ldots,P} \theta_p s_p(t)\right).$$

The mathematical form for the likelihood function for a sequence of events is equivalent to Cox’s (1972) proportional hazards model. In order to recover the rate parameters for a particular sequence of events, maximum likelihood estimation can be applied directly to the log-likelihood function. Alternatively, Bayesian estimation methods may also be used, and empirically have proven to be more efficient; for more detail, see Butts (2008).

14.3 Relational Event Models as an Assessment Tool

Evaluating process requires insight into the structure and evolution of team interactions over time. The encoding of structural signatures provides an unprecedented high-fidelity quantitative measure of the frequency with which certain behavioral patterns repeat themselves in an event history. As a result, the dynamics of team communication and collaboration can be explicitly studied at a resolution heretofore unavailable. Relational event models determine the relative influence of each SSS on future behaviors; this output is a standard statistical metric that can be compared across teams. By using SSS as a metric for analyzing team actions, outcomes can be explained as an explicit and direct result of the structure and nature of team process.

At the group or network level, SSSs represent the prevalence of certain behavioral patterns in an interaction network. Differences in the emergence of these mechanisms across teams or across individuals are indicative of structural variations in the interaction patterns of individuals and/or teams. The variability in the estimated values of REM parameters for different teams can be used to explain variability in the outcomes of these teams. To capture this impact, standardized relational event parameter estimates are used as independent variables in a statistical analysis where team outcomes such as performance or creativity are the dependent variable.

14.3.1 Example Using Relational Event Models as an Assessment Tool

To illustrate how relational event models are used to assess the effectiveness of multiple collaborative efforts, consider our previous example of individuals working on a software project. Suppose that our metric of interest is the SSS from Fig. 14.2, which measures the propensity for a team member to redo another member's work, based on their communication and the discrepancy in their skills. Let us assume that there are a number of these teams working on different software projects, and there is some measure of output quality, such as reliability from crashes or number of downloads by users, that can be compared across the software projects.

Using REM, we can estimate the parameter associated with our hypothesized SSS for each team. This output represents the degree to which each group engaged in that particular behavioral pattern during the course of their collaboration. We may compare these values across teams and determine the extent to which variation in behavior explains variations in output. This form of analysis allows us to answer the following question: "If a team more frequently engages in behavior X, will their collaboration result in a better output Y?"

14.4 Discussion

The study of effective collaborations requires an understanding of how individuals express their collaborative fluency, or collaboration skill. Unfortunately, measuring these processes has been a challenge. A gap exists between theories of effective collaboration (Olson, Malone, & Smith, 2001; Olson, Zimmerman, & Bos, 2008) and the methodological frameworks available to articulate and test those mechanisms; however, given the increased availability of digital trace data, complex interpersonal interactions are now made visible. The relational event framework is a statistical tool designed specifically to take advantage of this newly available data to make statistical inferences about what sequential patterns of collaboration ties form and how these patterns perform.

Previous methodologies typically focused on the nature or quality of aggregated interaction, without factoring in the rhythm, pattern, or tempo. Encoding individual actions and relations as temporal events can capture dynamic team processes with high levels of precision. SSSs, which are functions of event histories, represent dynamic interaction patterns that explain emergent behavior. These metrics are highly flexible and customizable to the context of the collaboration.

The relational event framework reveals behavioral patterns that can be used to assess the quality of a team's process with regard to the desired outcome of the collaboration. In general, the relational event methodology is geared towards understanding how teams work together, how teams communicate, and how they interact with the tasks and tools at hand. Relational event modeling is an exciting new statistical tool that allows for the development and testing of theory regarding the nature and quality of collaboration.

Acknowledgements The preparation of this chapter was supported by the Army Research Institute for the Social and Behavioral Sciences under contract W5J9CQ12C0017. The views, opinions, and/or findings contained in this report are those of the authors, and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

References

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.

Blossfeld, H.-P., & Rohwer, G. (1995). *Techniques of event history modelling: New approaches to causal analysis*. Mahwah, NJ: Erlbaum.

Brandes, U., Lerner, J., & Snijders, T. A. B. (2009, July). *Networks evolving step by step: Statistical analysis of dyadic event data.* Paper presented at the IEEE/ACM International Conference on Advances in Social Network Analysis and Mining, Athens, Greece.

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology, 38*(1), 155–200.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187–220.

Crawford, E., & LePine, J. (2013). A configural theory of team processes: Accounting for the structure of taskwork and teamwork. *Academy of Management Review, 38*(1), 32–48.

Foucault Welles, B., Vashevko, A., Bennett, N., & Contractor, N. (2014). Dynamic models of communication in an online friendship network. *Communication Methods and Measures, 8*(4), 223–243.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco, CA: Jossey-Bass.

Leenders, R., Contractor, N., & DeChurch, L. (2015). Once upon a time: Understanding team processes as relational event networks. *Organizational Psychology Review., 6*(1), 92–115.

Lusher, D., Koskinen, J., & Robins, G. (Eds.). (2012). *Exponential random graph models for social networks: Theory, methods, and applications*. New York, NY: Cambridge University Press.

Marcum, C. S., & Butts, C. T. (2015). Constructing and modifying sequence statistics for relevent using informR in R. *Journal of Statistical Software, 64*(5), 1–36.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review, 26*(3), 356–376.

Olson, G. M., Malone, T. W., & Smith, J. B. (Eds.). (2001). *Coordination theory and collaboration technology*. Mahwah, NJ: Erlbaum.

Olson, G. M., Zimmerman, A., & Bos, N. (Eds.). (2008). *Scientific collaboration on the Internet*. Cambridge, MA: MIT Press.

Quintane, E., Conaldi, G., Tonellato, M., & Lomi, A. (2014). Modeling relational events: A case study on an open source software project. *Organizational Research Methods, 17*(1), 23–50.

Snijders, T. A. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology, 21*(1–2), 149–172.

Vu, D., Pattison, P., & Robins, G. (2015). Relational event models for social learning in MOOCs. *Social Networks, 43,* 121–135.

Chapter 15
Modeling Collaboration Using Point Processes

Peter F. Halpin and Alina A. von Davier

Abstract In this chapter, we outline the uses of point processes and related methods for modeling temporal dependence in human interactions. We begin by describing our example, which was drawn from teamwork in sports. We then discuss three interrelated steps in analyzing the data: (a) the problem of defining and detecting temporal dependence among the activities of team members, (b) characterization of the dependence in terms of temporal clustering, and (c) the use of the Hawkes process to model the clustering. The third step provides a parametric model for describing and comparing statistical regularities of the interactions among individual team members or subsets of team members. We conclude by considering how this approach can capture aspects of team interaction that might be relevant for developing performance-based assessments involving collaborative problem solving.

Keywords Point processes · Hawkes processes · Human interaction · Collaboration

15.1 Introduction

The goal of this chapter is to provide an overview of some related methods for analyzing temporal dependence in human interactions. The overall modeling approach and its potential application to collaborative problem solving (CPS) have been discussed previously by von Davier and Halpin (2013). The contribution of

This work was conducted while Alina A. von Davier was employed with Educational Testing Service.

P.F. Halpin (✉)
New York University, New York, NY, USA
e-mail: peter.halpin@nyu.edu

A.A. von Davier
ACT, Iowa City, IA, USA
e-mail: Alina.vondavier@act.org

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_15

this chapter is to provide an accessible introduction, illustrated with an example from teamwork in sports.

As demonstrated by the breadth of contributions to this edited volume, CPS intersects with many domains of study. Within this context, the methodology presented in this chapter can be viewed as addressing the problem of how to quantify the contributions of individual team members, or subsets of team members, to their team's performance. Note that we use the term *performance* in a descriptive sense—to describe how team members interact while carrying out a task. Although team performance may be interpreted also to include an evaluative component (e.g., whether a task was completed successfully), this is not our focus here. We also do not address the problem of defining or directly measuring the skills that might make an individual a good collaborator (e.g., Griffin & Care, 2015; Griffin, McGaw, & Care, 2012; Hao, von Davier, & Kyllonen, 2015; Liu, Hao, von Davier, & Kyllonen, 2015). However, by showing how existing statistical methods can be used to examine team interactions, we hope that the present research can contribute to the study of such skills in performance-based settings.

Section 15.2 introduces our example, which we draw on throughout the chapter to illustrate the methodology. The data were taken from a single professional basketball team playing a single game: the Philadelphia 76ers in game four of their 1984 playoff appearance against the Chicago Bulls. While the example provides an intuitive context to study temporal dependence among the actions of team members, it also has several limitations that make the comparison with CPS in educational and assessment settings somewhat strained (e.g., the role of team training; player substitutions during game play). Therefore we defer explicit consideration of how these methods might be applied in assessment settings until the concluding section of this chapter.

In Sect. 15.3, we introduce the concept of an event time, which is the basic unit of analysis of the methods we propose. In the context of teamwork, an event can represent any human action that has negligible duration, relative to the period of observation under consideration. Events can be contrasted with states or regimes, which persist in time. In our example, the observation period under consideration is the time-on-offensive of the 76ers, and the events we focus on are (a) passes and (b) shots on basket.

In Sect. 15.3, we also present a quantitative definition of temporal dependence for event times in terms of time-lagged mutual information (e.g., Brillinger, 2004; Cover & Thomas, 2005). Importantly, this definition can be applied to any subset of a team's members, including the individual members themselves. We use the term *team unit* to refer to a subset of interest. When considering a single team unit in isolation, time-lagged mutual information describes the dependence of that unit's actions on its own past actions, which we refer to as *intradependence*. When considering dependence between team units, we use the term *interdependence*. In general, we suggest that intradependence and interdependence provide two competing explanations of team performance. We also use our example to show how mutual information can be used as a data-analytic method for inferring the presence of either type of dependence.

Having defined temporal dependence for event times, the second step in our analytic approach is to characterize the nature of the dependence. Past research on human dynamics supports the hypothesis that many types of human interaction are temporally clustered (e.g., Barabási, 2005; Crane & Sornette, 2008; Halpin & De Boeck, 2013; Matsubara, Sakurai, Prakash, Li, & Faloutsos, 2012; Oliveira & Vazquez, 2009). In the context of team interdependence, clustering means that the actions of one team unit are associated with increased probability of further actions by other units in the near future. Referring again to our example, this means that Team Unit A passing the ball to Team Unit B should be associated with increased probability that Team Unit B will pass or shoot the ball in the near future. In Sect. 15.4, we discuss methods to assess temporal clustering in event data, and, unsurprisingly, we find that such a pattern is present in our example.

In the penultimate section, we describe the third step, which is to develop a parametric model for the clustered event times. There is a general divide between models used for clustered data (e.g., excitatory processes) and those used for nonclustered data (e.g., regulatory processes). The Hawkes process (Hawkes, 1971; Hawkes & Oakes, 1974) provides a relatively general framework for clustered event times. The model analyzes the overall temporal dependence of a team's actions in terms of statistical regularities in the responsiveness of the team units to themselves and to one another. As illustrated in our example, the parameters of the model allow us to characterize and compare the performance of team units and to describe overall dynamics of the team interactions.

In the final section, we summarize the methodology and analysis with an emphasis on potential applications to performance-based assessments of CPS, and we discuss limitations and future research directions.

15.2 Description of Example Data

As mentioned, our example is taken from a single game of professional basketball. Some reasons for using professional basketball to illustrate the application of point process to CPS are that (a) data are publicly available, (b) many of the "moves" made by individual team members satisfy the requirement of having negligible duration, and (c) it is intuitive that the moves of one player can depend on those of the other team members. Although play-by-play data sets are available for purchase from commercial vendors,[1] these data sets only record moves that are directly associated with scoring, mainly shots on basket and rebounds. This limits the

[1] For example, http://www.basketballgeek.com/data/.

opportunity to examine team dynamics within an offensive possession because no information about passes is recorded. To obtain the present data set, which also included passes, we manually coded video data. This allowed for a more nuanced examination of team interaction but limited the duration of team play we were able to consider. In general terms, an ideal data set would encode a wide range of task-related activities over an extended period of time.

In the game we analyzed, the 76ers had a total of 92 offensive possessions resulting in a total of $N = 401$ passes and shots. We can break down the events in terms of a factorial design: Player × Type (pass or shot) × Success (whether a pass was completed or a shot resulted in a score). If the same five players were on the court for the entire game, this would result in 20 cells. However, player substitutions were made throughout the game, with a total of nine players spending time on the court (= 36 cells). Many of these cells are empty or have very low event counts, and consequently, there is not sufficient data in the single game for time series analysis at the cell level. Therefore we make some necessary compromises in our data analyses.

Because our main concern is team interaction, we focus on the marginal processes for players, ignoring the type and success of each event. To deal with substitutions and small event counts for players who were infrequently on the court, we divided the team into three team units: (a) the point guards (Players 1 and 12), one and only one of whom was always on the court and who had the role of bringing the ball up the court and setting up the offense; (b) Charles Barkley (Player 34), who was the star player of the 76ers team, played most of the game, and had a unique role as power forward and main scorer of the evening; and (c) other players, or all other members of the team. Placing the rest of the team under one label meant that we could ignore player substitutions, which greatly simplified the analysis. However, with a larger data set, the other players could be further divided into more meaningful team units, for example, the other team positions (e.g., shooting guard, center, small forward).

To focus our analysis on the 76ers' offense, the Bulls possessions were replaced with a short (≈7 s) random time buffer. This ensured that there was no statistical dependence between the 76ers' moves on subsequent possessions. Each time buffer was obtained as a random draw from an exponential distribution with rate parameter equal to the 95th percentile of the 76ers' waiting time distribution. We also omitted periods of time when Barkley was not on the court (≈3 min) as well as half time and TV commercials. All other breaks in play were retained (e.g., time-outs, free throws). The jump ball was used as $t = 0$, and time was recorded in seconds until the final minute of game time, after which the play consisted mainly of time-outs and free throws. Note that the recorded time does not denote the time on the game clock but rather the full duration of the 76ers' possessions, including time-outs and free throws. The resulting data are summarized in Fig. 15.1.

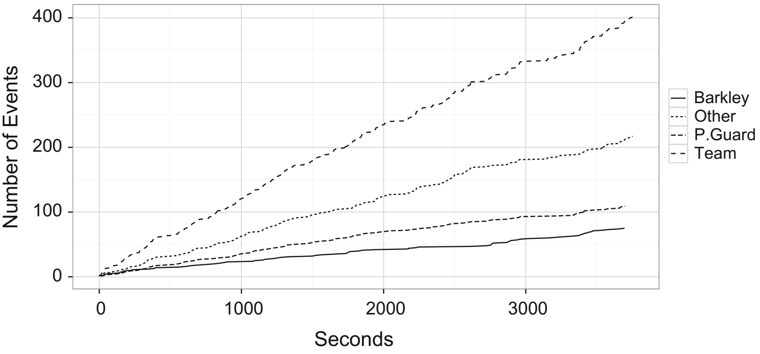


Fig. 15.1 Cumulative event count as a function of time for each player unit and the entire team

15.3 Defining Temporal Dependence

As noted earlier, the first challenge is to define whether, and to what extent, event times demonstrate temporal dependence.

To this end, let $X = (X_1, X_2, \ldots, X_N)$ and $Y = (Y_1, Y_2, \ldots, Y_M)$ be random variables denoting two sequences of event times. For example, the event times for Barkley begin (14.00, 19.33, 44.08), which denotes the time, in seconds, of his first three moves. Graphically, event times are the time points at which the cumulative event counts in Fig. 15.1 increase. Event times are also related to the familiar concept of waiting times or response times. In the univariate case, event times and waiting times provide equivalent representations of a point process (see, e.g., Daley & Vera-Jones, 2003); however, in the multivariate case, multiple definitions of a waiting time exist (i.e., between and within margins), and therefore the representation in terms of event times is more readily generalized.

The mutual information of X and Y is defined as

$$I_{XY} = E_{X,Y} \ln\left[\frac{f(X,Y)}{f_X(X) f_Y(Y)}\right], \tag{15.1}$$

where f is the joint probability density function, f_X and f_Y are the marginals, E_U denotes expectation over the distribution of U, and ln is the natural logarithm.

An accessible discussion of the theoretical and data-analytic underpinnings of mutual information is given in Brillinger (2004). Here we simply list some of its more useful characteristics:

1. $I_{XY} = 0$ if and only if X and Y are statistically independent; otherwise, $I_{XY} > 0$.
2. I_{XY} makes mild assumptions about the kind of relationship between X and Y; in particular, the relationship can be nonlinear.

3. When $I_{XY} = 0$, its sample estimate (see Eq. 15.2) has a known sampling distribution; this provides confidence bounds on the hypothesis of "no team interaction."
4. I_{XY} can be generalized to more than two sequences of event times. For example, consider three sequences X, Y_1, Y_2. Then define $Y = Y_1 + Y_2$, and apply the definition of Eq. 15.1 as earlier. Here the plus sign denotes "superposition" of Y_1 and Y_2, which is just the sequence of all the event times in Y_1 and Y_2 (for a technical discussion, see Daley & Vera-Jones, 2003). Although the choice of X and Y will depend on the specific questions at hand, we have found it useful to let X denote a single team unit and Y denote all remaining team units. This allows us to describe how the actions of any one team unit depend on those of the rest of the team.
5. I_{XY} readily incorporates historical dependence, which we have referred to previously as time-lagged mutual information. For example, let $Y = X - a$, where $a > 0$ denotes a fixed constant that is subtracted from each element of X. Then $I_{XY} = I_{XX-a}$ denotes the dependence of an event stream on its own past, at lag a. In the introduction, we referred to this as intradependence. Similarly, I_{XY-a} denotes the dependence of X on the past of Y, at lag a. This is what we have referred to in the introduction as interdependence. As illustrated subsequently, treating intra- and interdependence as a function of a allows for a description of how the timing of events depends on the timing of past events.

A rough "plug-in" sample estimate of I_{XY} for event data can be obtained as follows (for a review of other approaches, see Paninski, 2003). First, discretize the time interval $[0, T]$, for example, into $k = 1, \ldots, K$ bins of size $\delta = T/K$. Here T denotes the end time of the observation period. For our example data, $T \approx 3744$ s, or about 62 min. Next, recode the event stream X in terms of K realizations of a random variable U defined such that $u_k = 1$ if an X-event happens in the interval $z_k = [\delta(k-1),\ \delta k)$ and $u_k = 0$ otherwise. Similarly, recode Y using the random variable W with realizations w_k. Then

$$I_{XY} \approx I_{UW} = \sum_{i,j \in \{0,1\}} p_{ij} \ln\left(\frac{p_{ij}}{p_{i+}\, p_{+j}}\right), \tag{15.2}$$

where $p_{ij} = \text{Prob}\{U = i, W = j\}$ and p_{i+} and p_{+j} are the marginals. When U and W are independent, the sampling distribution of I_{UW} is proportional to that of a chi-square statistic on 1 degree of freedom (see Brillinger, 2004).

There are two main limitations of this approach to estimation. The first is that more than one event may fall into a single interval z_k, in which case U and W are not good approximations to X and Y. In theory, this is not a severe problem, because the coarseness of the discretization is under the control of the analyst, and as $K \to \infty$, the approximation becomes exact. However, computational time depends on the value of K, so it is usually preferable to use a relatively small number of intervals. In our example, we chose $\delta = 1$ s for the bin width, which implied $K = 3374$ bins.

The second limitation concerns the estimation of p_{ij}. In practice, we have used the standard maximum likelihood estimate for i.i.d. data, $\hat{p}_{ij} = \sum_k u_k w_k / K$, which is unbiased when $I_{UW} = 0$ but is otherwise motivated only by its convenience. In general, addressing serial dependence in binary data requires specification of a model for the dependence (e.g., Budescu, 1985). We reserve the use of an explicit model until Sect. 15.5, where we introduce the Hawkes process.

Despite these limitations, we have found that Eq. 15.2 can be a useful tool for data mining—it makes weak assumptions about the statistical nature of the signal and is scalable to very large data sets, because its computational complexity grows with K rather than with the number of events. Conversely, the estimation of a parametric model such as the Hawkes process is sensitive to model misspecification and relatively computationally demanding. We therefore recommend using Eq. 15.2 for making a "first pass" at the data.

Figure 15.2 depicts the sample intra- and interdependence functions for the example data, treated as a function of time lag $a \in [0, 20]$ s. The left panel shows how the moves of each team unit depend on the unit's own past moves. The right panel shows how the moves of each team unit depend on the past moves of all remaining team units. Both panels also depict how the moves of the entire team depend on the past of the entire team (i.e., the functions for "team" are the same in both panels). The shaded areas denote the 99% confidence interval on the null hypothesis that $I_{UW} = 0$.

Examining the team dependence functions, we see that the offensive moves of the entire team depend on their past moves within a window of about 5 s. This means that the probability of any player on the team shooting or passing the ball at any given point in time depended on the moves that the team had made within the last 5 s. However, as shown in the left panel, none of the individual team units demonstrated intradependence. Consequently, the dependence at the "team level" cannot be explained in terms of the moves of any single team unit considered in

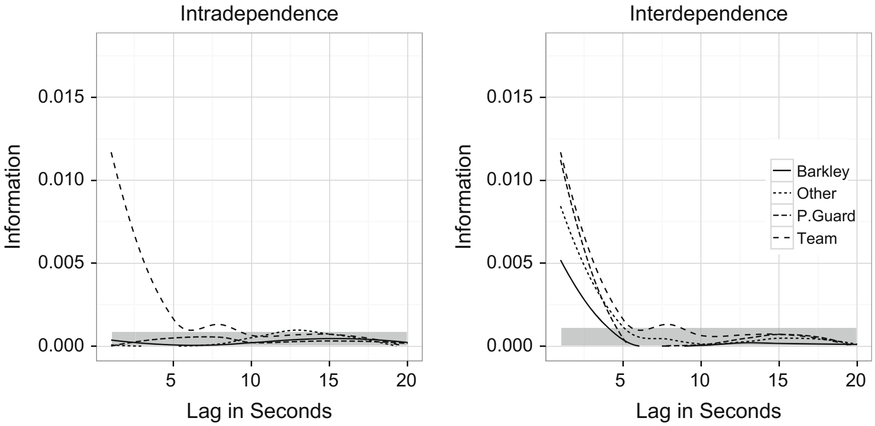


Fig. 15.2 Sample intra- and interdependence functions for the entire team and the player units, as a function of time lag

isolation. By contrast, the right panel shows that the moves of each team unit depended on those of the rest of the team. In other words, the probability that each player will make a move at a given point in time depended on recent moves of the other team units. Thus we can characterize the 76ers' offense as demonstrating strong interdependence and weak intradependence among the three team units.

In the context of a basketball game, these findings should not be surprising: whether I shoot or pass the ball depends on whether another player has passed the ball to me. In the conclusion, we discuss the utility of inferring such patterns when the dependence structure is not so obvious.

15.4 Assessment of Clustering

Having inferred temporal interdependence among the actions of the team units, the next step in our approach is to characterize the nature of that dependence. In particular, we assess the number of events occurring during an observation period for overdispersion relative to the Poisson distribution. If the point process generating the event times is a homogeneous Poisson process, then the number of points occurring in the observation period has a Poisson distribution. On the other hand, if the number of points is overdispersed relative to the Poisson distribution, then the event times occur in "clusters" of relatively high frequency, separated by periods of relatively low frequency (see Daley & Vera-Jones, 2003). For this reason, such data are commonly referred to as clustered. Synthetic examples of clustered and unclustered data are shown in Fig. 15.3, where $N \approx 300$ for both examples.

As noted, there is empirical evidence to support the hypothesis that clustering is a prevalent characteristic of human interaction (e.g., Barabási, 2005; Crane & Sornette, 2008; Halpin & De Boeck, 2013; Matsubara et al., 2012; Oliveira & Vazquez, 2009). In addition to the terminology of clustering and overdispersion, this research has often been phrased in terms of *bursts*, *heavy-tailed waiting time distributions*, or *power law distributions*. The hypothesis of clustering also seems plausible in the context of teamwork and collaboration. Here clustering would mean that the actions of one team unit increase the probability of future actions by other units, and vice versa. Alternatively, if the actions of a team unit consistently retard further actions of the team, this may seem antithetical to team work in many contexts. However, it remains an empirical question whether clustering is useful for characterizing teamwork in general.

To test for clustering of the basketball data, we used an approach motivated by the time-change theorem (see, e.g., Daley & Vera-Jones, 2003, Chap. 7). The theorem states that the waiting times of the residuals of a correctly specified point process are exponentially distributed with a rate of 1. This result provides a relatively general approach for assessing the goodness of fit of point process models.

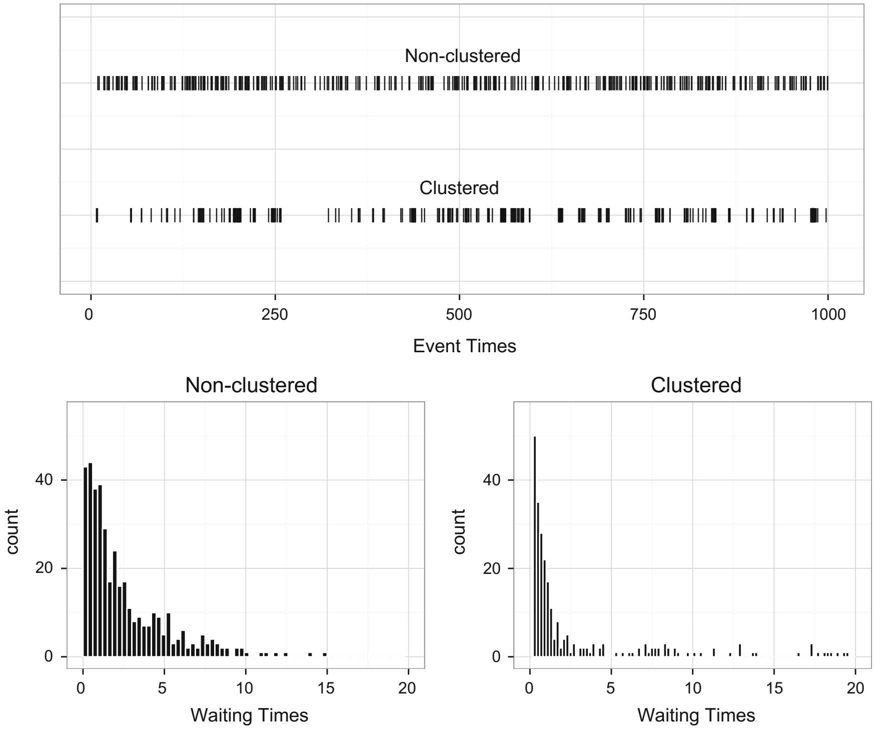


Fig. 15.3 Event times and waiting times for clustered and nonclustered data (synthetic example)

To test for overdispersion, we fitted the homogeneous Poisson process to the event times for the full team and to the waiting times for each team unit (waiting time was defined as the time since the previous team event, not the previous event of that team unit). The residual analysis is summarized in Fig. 15.4.

It is important to note that we have already provided an initial assessment of whether the basketball data are compatible with a homogeneous Poisson process, because in that case, their interdependence functions (Fig. 15.2, right) would be equal to zero at all lags. However, here we are looking for a specific pattern of deviation from the Poisson model, namely, clustering. In the QQ plots, clustering is evidenced by an S-shaped pattern around the reference line, where we see more short waiting times than expected and also many waiting times that are longer than expected. This pattern is apparent for the overall team as well as for each team unit. Thus we conclude that the interdependence exhibited by the basketball example (Fig. 15.2) is characterized by clustering (Fig. 15.4).

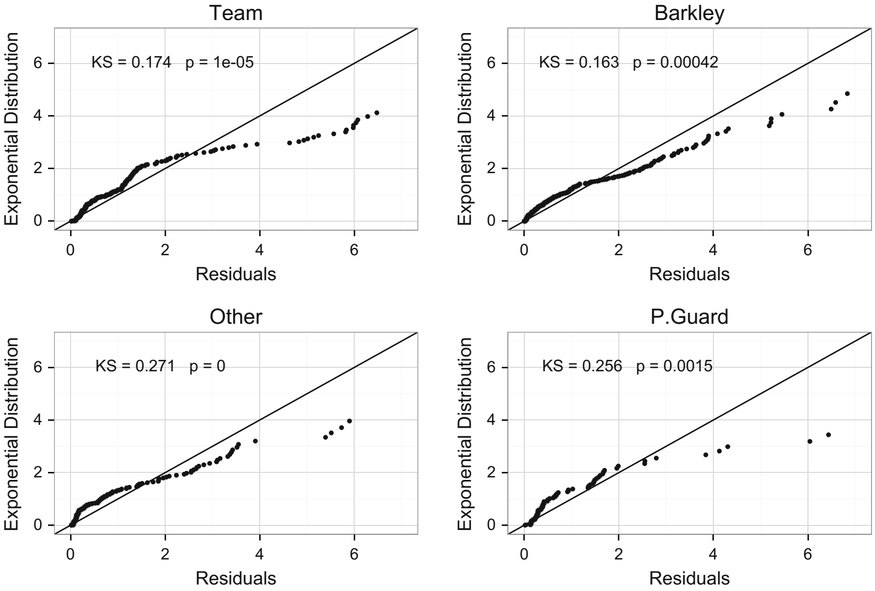


Fig. 15.4 Quantile-quantile plots of residual waiting times from the Poisson process against the exponential distribution with rate of 1. *Insets* give Kolmogorov–Smirnoff test and its two-sided *p*-value

15.5 Modeling Team Interactions Using the Hawkes Process

At this point, we have provided evidence that our example data exhibited temporal dependence and that this dependence was due to clustering. An appropriate statistical model for clustered event times is the Hawkes process (e.g., Hawkes, 1971; Hawkes & Oakes, 1974). In the context of our example, the Hawkes process describes each team unit's actions as responses to its own past actions or as responses to the actions of the other team units. The overall hypothesis of the model is that this responsiveness accounts for the observed clustering—a hypothesis that is intuitively obvious in the case of basketball.

Owing to space constraints, we do not cover the formal specification of the Hawkes process, which is available from many other sources (e.g., Brillinger, 1975; Daley & Vera-Jones, 2003; Halpin & De Boeck, 2013; Hawkes, 1971; Rasmussen, 2012). Instead, we refer to Eq. 15.3 to provide a nontechnical explanation of the overall setup of the model:

$$\Phi(t) = \begin{bmatrix} \phi_{BB}(t) & \phi_{BO}(t) & \phi_{BP}(t) \\ \phi_{OB}(t) & \phi_{OO}(t) & \phi_{OP}(t) \\ \phi_{PB}(t) & \phi_{PO}(t) & \phi_{PP}(t) \end{bmatrix}. \tag{15.3}$$

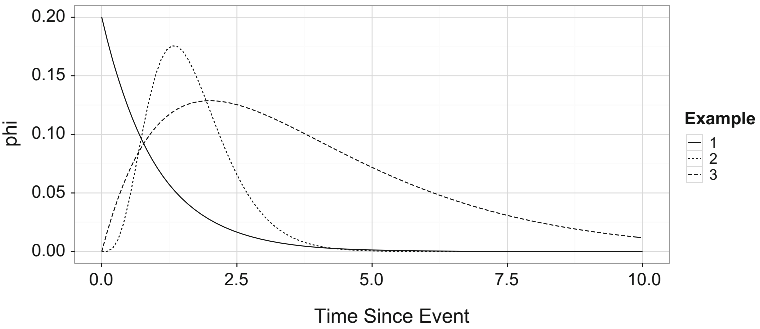


Fig. 15.5 Three examples of response functions using two-parameter gamma density

Here $\Phi(t)$ is a matrix of response functions, each denoted $\phi_{ij}(t)$. The subscripts stand for team units: B = Barkley, O = other members, and P = point guard. The first subscript indexes the team unit that is responding (the *output process*), and the second indexes the team unit to which the response is made (the *input process*). The diagonal elements of $\Phi(t)$ represent the responsiveness of a team unit to its own past actions. As per our initial analyses (Sect. 15.3), we set the diagonal elements to zero—there was no evidence of intradependence in the present data. The off-diagonal elements represents the responsiveness of a team unit to the actions of the other team units. Because we have found evidence of interdependence in the data, these response functions are the focus of the present analysis.

For the Hawkes process, the response functions may be written as $\phi_{ij}(t) = \alpha_{ij} \times f(t; \xi_{ij})$, where $\alpha_{ij} \in (0, 1)$ is called the intensity parameter and f is a probability density function defined on $\mathbb{R}+$ with parameter vector ξ_{ij}. For this analysis, we let f be the two-parameter gamma density, with some example response functions depicted in Fig. 15.5. Referring to the figure, the value of zero on the horizontal axis represents the time of an event, and as we travel down the horizontal axis, that event recedes into the past. The role of the response function is to describe how the event is associated with the probability of another event occurring in the near future. Otherwise stated, the response curves characterize the "memory" of the different team units.

We estimated the Hawkes process for the basketball data using the expectation-maximization algorithm developed in Halpin and De Boeck (2013) and Halpin (2013). The goodness of fit of the model was assessed using the same approach discussed for the Poisson process in Fig. 15.4. The residuals for the Hawkes process are shown in Fig. 15.6; overall, they exhibit agreement with their hypothetical distribution. For the point guards, it appears that the Hawkes process may have overcorrected the clustering to some degree, but the Kolmogorov–Smirnoff test was not significant. We conclude that the Hawkes process adequately accounted for the clustering in the data.

Moving on to interpret the model parameters, Fig. 15.7 depicts the estimated response functions, $\hat{\phi}_{ij}(t) = \hat{\alpha}_{ij} \times f(t; \hat{\xi}_{ij})$, corresponding to the off-diagonal

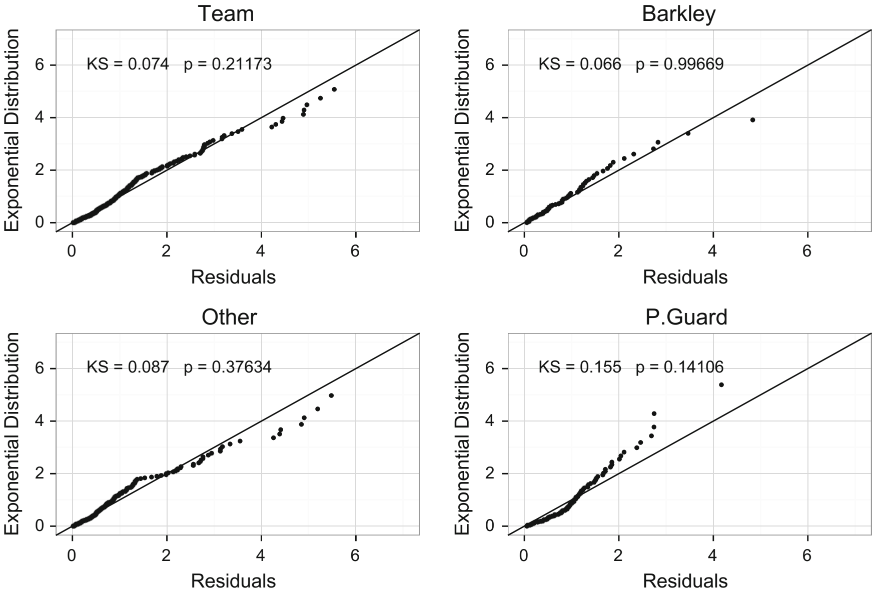


Fig. 15.6 Quantile-quantile plots of residual waiting times from the Hawkes process against the exponential distribution with rate of 1. *Insets* give Kolmogorov–Smirnoff test and its two-sided *p*-value

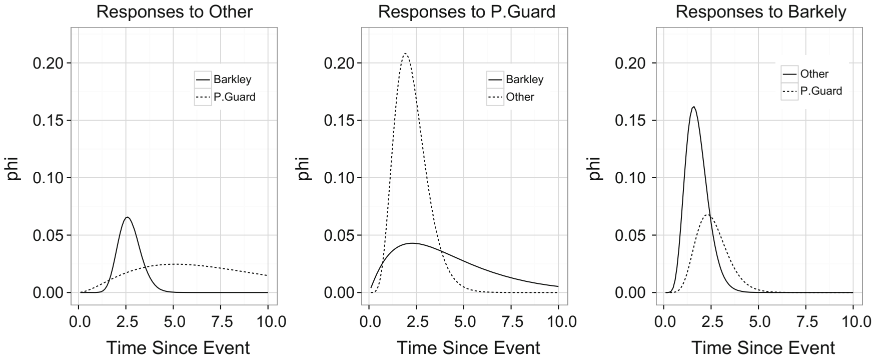


Fig. 15.7 Estimated response functions for the basketball data

elements of Eq. 15.3. Considering the leftmost panel, we may conclude that Barkley's moves depended on those of the other members during a window of about 2–4 s lag. This is plausible, because Barkley was usually positioned under the net, so that once he received the ball, he was likely to shoot or pass right away. In contrast, the point guards' dependence on the other members was more protracted.

This is explainable by the fact that the point guard is responsible for bringing the ball up the court and setting up the offense after receiving an inbound pass. The other plots can be interpreted in a similar manner.

It is also possible to test for equalities between the different response functions or whether the parameters of the model are equal to prespecified values of interest using standard procedures for maximum likelihood (e.g., the Wald test). In particular, many hypotheses about overall team dynamics can be formulated in terms of the matrix $\Phi(t)$. For example, if the matrix were inferred to be symmetrical, we might consider this evidence of reciprocity among team units. Alternatively, a team could demonstrate reciprocity in its overall responsiveness (i.e., in α_{ij}) but not in the shape or memory of the response function. Or perhaps only some team units demonstrate equality of responsiveness and others do not. Clearly many hypotheses about team dynamics can be formulated and tested, depending on the research questions and the interests of the analyst.

15.6 Future Directions and Limitations in Assessment Applications

We hope that the foregoing sections have convinced the reader that point processes can provide an interesting perspective on human interactions and teamwork. We also hope that our suggestions about how to carry out such analyses are a useful starting point for further research on this topic. However, we have not yet made any explicit connections with the assessment context, which is the purpose of this concluding section. In general, we suggest that the methods discussed here provide a viable means of analyzing process data obtained from performance-based assessments of CPS. However, such assessments are in fact not widely available, for which reason we have used an example from sports rather than from the assessment literature. Therefore, in considering future directions and limitations, we are essentially imagining how assessments of CPS could be designed to take advantage of point process methodology and in what ways the methodology may fail to answer important questions about CPS. Naturally, these final considerations are not intended to be exhaustive but merely to address some points we find most salient.

15.6.1 Designing Tasks to Measure Team Interactions

By analyzing the interactions of real teams in terms of concepts such as team units, intra- and interdependence, temporal clustering, and reciprocity, we can make some headway on assessing how team members interact to achieve their goals. This perspective invites the development of performance-based assessments that can provide evidence about the presence or absence of such team dynamics. For example, if we wanted to assess whether a team was able to demonstrate higher levels of interdependence than intradependence, what types of tasks would we design? If we wanted to know whether the actions of one team unit were associated with increased probability of further actions by other team units, what types of tasks would we design? These are currently wide open questions in the assessment literature. Yet, if we desire to move beyond self-reported measures of CPS, we surely require performance-based contexts that allow for the interactions of collaborators to be exhibited and recorded.

15.6.2 Modeling the Relation Between Team Interactions and Individual Ability

In terms of our basketball example, each shot on basket can be successful or not, and the overall success of a player in making shots can be interpreted as evidence about his or her ability as a basketball player. This general idea is similar to how psychological and educational test are scored, and we elaborate on the basketball example with this analogy in mind. On the basis of the analysis reported in Sect. 15.3, we may conclude that the timing of a shot can be predicted by the timing of passes leading up to the shot, within a window of about 5 s. However, we have not addressed the relation between the *success* of a shot and the timing of the passes leading up to it. It may be the case that the success of a shot is simply not predictable from the team dynamics. Alternatively, by using the sequence and/or timing of events leading up to a shot, we might gain a much better understanding of its likelihood of success. This point has not been addressed by the methodology discussed in this chapter, and this is a major shortcoming in establishing the utility of proposed methods for supporting inferences about the ability of individual team members. We anticipate that progress on this front can be made by application of marked point processes (e.g., Daley & Vera-Jones, 2003), wherein the successful completion of a task component can be modeled as a time-varying covariate within the point process framework. The interested reader is referred to von Davier and Halpin (2013) for further discussion.

References

Barabási, A. L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature, 435*, 207–211. doi:10.1038/nature03526.1

Brillinger, D. R. (1975). The identification of point process systems. *Annals of Probability, 3*, 909–929.

Brillinger, D. R. (2004). Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics, 18*, 163–182.

Budescu, D. V. (1985). Analysis of dichotomous variables in the presence of serial dependence. *Psychological Bulletin, 97*, 547–561.

Cover, T. M., & Thomas, J. A. (2005). *Elements of information theory*. New York, NY, USA: Wiley.

Crane, R., & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 15649–15653.

Daley, D. J., & Vera-Jones, D. (2003). *An introduction to the theory of point processes: Elementary theory and methods* (2nd ed., Vol. 1). New York, NY, USA: Springer.

Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century skills: Methods and approach*. New York, NY, USA: Springer.

Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills*. New York, NY, USA: Springer.

Halpin, P. F., & De Boeck, P. (2013). Modelling dyadic interaction with Hawkes processes. *Psychometrika, 78*, 793–814. doi:10.1007/s11336-013-9329-1

Hao, J., Liu, L., von Davier, A., & Kyllonen, P. (2015). Assessing collaborative problem solving with simulation-based tasks. In O. Lindwall, P. Häkkine, T. Koschmann, P. Tchounikine, & S. Ludvigsen (Eds.), *Exploring the material conditions of learning: The computer supported collaborative learning conference* (Vol. 2, pp. 544–547). Gothenberg, Sweden: International Society of the Learning Sciences.

Halpin, P. F. (2013). A scalable EM algorithm for Hawkes processes. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Proceedings of the 77th international meeting of the psychometric society* (pp. 403–414). New York: Springer.

Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society, Series B33, 104*, 438–443. doi:10.1073/pnas.0703993104

Hawkes, A. G., & Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability, 11*, 493–503.

Liu, L., Hao, J., von Davier, A., Kyllonen, P., & Zapata-Rivera, D. (2015). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on computational tools for real-world skill development* (pp. 344–359). Hershey, PA, USA: IGI-Global.

Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L., & Faloutsos, C. (2012). Rise and fall patterns of information diffusion: Model and implications. In *KDD'12: Proceedings of the 18th ACM SIGKDD* (pp. 6–14). New York, NY, USA: ACM.

Oliveira, J. G., & Vazquez, A. (2009). Impact of interactions on human dynamics Impact of interactions on human dynamics. *Physica A, 388*, 187–192.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation, 15*, 1191–1253. doi:10.1162/089976603321780272

Rasmussen, J. G. (2012). Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability, 15*, 623–642. doi:10.1007/s11009-011-9272-5

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations* (ETS Research Report No. RR-13-41). Princeton, NJ, USA: ETS.

Chapter 16
Dynamic Bayesian Network Models for Peer Tutoring Interactions

Yoav Bergner, Erin Walker, and Amy Ogan

Abstract The ability to automatically distinguish between effective and ineffective patterns in collaborative learning sessions opens doors to improved opportunity for learning in pairs or groups even when a teacher might not be available to facilitate. In this chapter, data from one-time computer-based peer tutoring sessions are modeled using hidden Markov models (HMMs) in two ways. The first model uses an input–output HMM to compare the assistance value of different tutor inputs in helping the tutee correct a mistaken step in solution. This model uses only automatically generated codes based on context and cognitive content of the tutor chat. The second model predicts tutee normalized gains from pre- to posttest in the experimental condition. Both cognitive and affective labels to tutor chats (human coded) were included as well as tutee (in)correctness, undos, and chats back to the tutor. Performance of the HMM is favorable compared to a "static" logistic regression model using aggregated totals of the same observables. Some of the hidden states are readily interpretable, though deeper comparison between high- and low-gain groups is part of ongoing work.

Keywords Bayesian networks · Hidden markov models · Tutoring · Collaboration

This work was conducted while Yoav Bergner was employed with Educational Testing Service.

Y. Bergner (✉)
New York University, New York, USA
e-mail: yoav.bergner@nyu.edu

E. Walker
Arizona State University, Tempe, AZ, USA
e-mail: erin.a.walker@asu.edu

A. Ogan
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: aeo@cs.cmu.edu

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_16

16.1 Introduction

An important aspect of learning is the social construction of knowledge, where students exchange ideas, reflect on their own misconceptions, and come to a shared understanding through dialogue with their peers (Schoenfeld, 1992). Unfortunately, students do not always engage in these positive interactions spontaneously. In recent years, there has been substantial interest in developing intelligent adaptive collaborative learning support (for a review, see Magnisalis, Demetriadis, & Karakostas, 2011). Students need some type of support—scripting, training, facilitation—to collaborate effectively. The promise of adaptive intelligent facilitation is that support is tailored directly to the individuals collaborating but is less resource intensive than support delivered by a human facilitator. Studies comparing adaptive support to static and no support conditions have indeed provided evidence that adaptive support might be an effective approach (Baghaei, Mitrovic, & Irwin, 2007; Kumar, Rosé, Wang, Joshi, & Robinson, 2007; Walker, Rummel, & Koedinger, 2014). A key component of adaptive support is modeling productive and unproductive collaborations (Soller & Stevens, 2007). Thus the ability to automatically distinguish between effective and ineffective patterns in collaborative learning opens doors to improved opportunity for learning in pairs or groups. It enables accurate prediction of collaborative benefit, so that a teacher or automated system may intervene where appropriate. It also facilitates the development of a deeper understanding of why certain collaborations are beneficial and others are less productive.

One potential approach for modeling sequential processes in collaborative learning is the use of hidden Markov models (HMMs). Soller and Stevens (2007) applied HMMs in modeling effective and ineffective knowledge-sharing sequences or problem-solving strategies. Observable states were determined by the human agent roles and constrained sentence openers, which were organized into deliberate categories (i.e., Inform, Request, or Acknowledge). Although results were promising in these studies, performance above baseline may have been exaggerated, because the sample had twice as many ineffective groups as effective groups. Boyer, Phillips, Ingram, and Ha (2011) applied HMMs to expert tutor behaviors to model effective tutoring strategies with undergraduate computer science students. Bigram analysis was used to determine adjacency pairs of actions. Thus annotated dialogue acts, task actions, or joined pairs of these constituted the set of observable actions, and a HMM classifier was trained on the data to distinguish between two different tutor styles. Learning gain itself was not predicted using a dynamic model, and only frequency counts of hidden states were used after the fact as predictors of learning gain from the session. Formally related work with dynamic models has included the use of partially observed Markov decision processes for student learning (Almond, 2007). Though not peer-to-peer collaborative, these models include decision variables for an activity chosen by an instructor and a hidden state representing a continuous measure of individual student proficiency.

In this chapter, we investigate two different approaches to modeling collaborative learning using dynamic Bayesian networks (DBNs). In the first application, we report on a model to compare the short-term assistance value of individual tutor inputs using an input–output hidden Markov model (IOHMM; Bengio & Frasconi, 1995). Assistance was operationalized as helping the tutee correct an error that he or she had made during problem solving. In the second application, we take a more holistic view of the tutor–tutee dyad and model the sequence of observed outputs from both using a single output layer, quite similar to the methodology of Soller and Stevens (2007). Both cognitive and affective labels to tutor chats (human coded) were included. Performance of the HMM is also compared to a "static" logistic regression model using aggregates of the same observables. By employing these two approaches, we demonstrate the viability of the use of HMMs in this type of modeling; furthermore, we improve understanding of patterns of positive and negative peer tutoring interactions. An improved understanding of help giving should apply to other collaborative scenarios where participants share information and give and receive help. Most elements of this dynamic modeling approach generalize to other types of collaboration as well.

The organization of this chapter is as follows. In Sect. 16.2, we provide some background about HMMs and the notational conventions of DBNs. In Sect. 16.3, we describe the Adaptive Peer Tutoring Assistant (APTA) and the nature of the data that were used for analysis. In Sects. 16.4 and 16.5, we describe the design of and results from applications of two different DBN models to these data. Conclusions and a discussion of future work follow.

16.2 Hidden Markov Models and Dynamic Bayesian Networks

In this section, we present the background on HMMs necessary for understanding the modeling techniques we use later in the chapter. The essence of HMMs can be traced at least to the work of Blackwell and Koopmans (1957) and Gilbert (1959), who considered the identifiability problem for functions of finite Markov chains. Baum and Petrie (1966) introduced a maximum likelihood estimation algorithm for HMM parameters well before the graphical model framework generalized them into the class of DBNs (Pearl, 1988). HMMs gained widespread use in signal processing applications, for example, automated speech (Rabiner, 1989), handwriting (Nag, Wong, & Fallside, 1986), and even sign language (Starner & Pentland, 1997) recognition. Soller and Stevens (2007) used more traditional HMM notation, whereas contemporary Bayesian knowledge tracing literature tends to use the DBN formalism (Reye, 2004). We adopt the modern DBN notation whereby a HMM is represented using a two-time-slice view, as shown in Fig. 16.1.

The edges in Fig. 16.1 explicitly denote conditional independence relationships. The Markov property is embedded in the factorization of the joint distribution for a

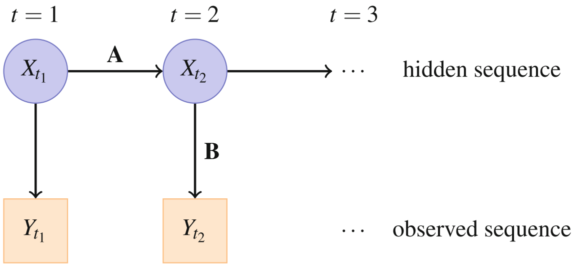


Fig. 16.1 Two-time-slice representation of a hidden Markov model

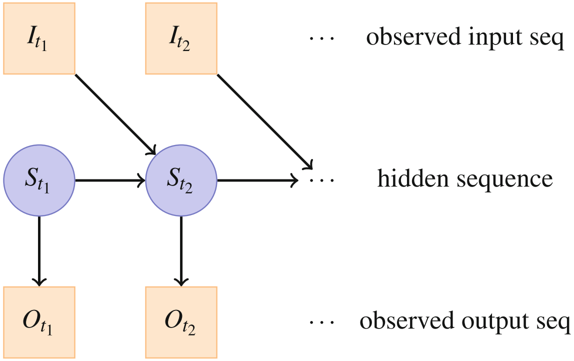


Fig. 16.2 An input–output hidden Markov model

given sequence of length T. Denoting the observed and hidden sequences by $Y_{1:T},\ X_{1:T}$, this joint distribution is compactly written (Ghahramani, 2001) as

$$P(X_{1:T},\ Y_{1:T}) = P(X_1)P(Y_1|X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})P(Y_t|X_t).$$

Older conventions (Baum & Petrie, 1966) denote the HMM by the set λ of prior probabilities π, (hidden) transitions **A**, and emissions **B**. We have added edge labels for the transition and emission matrices to Fig. 16.2. The connection is as follows:

$$\begin{aligned} \pi &\Leftrightarrow P(X_1), \\ \mathbf{A} &\Leftrightarrow P(X_t|X_{t-1}), \\ \mathbf{B} &\Leftrightarrow P(Y_t|X_t). \end{aligned}$$

Consider further the IOHMM (Bengio & Frasconi, 1995) in Fig. 16.2. The added layer of observed nodes now represents a second observed sequence, one which is understood to affect the output sequence through the mediation of a hidden state.

Alternately, the hidden state at each time slice depends not only on the previous hidden state but also on the preceding input. Enumerating the possible input states and hidden states by $I_t \in \{1 \ldots K\}$, $S_t \in \{1 \ldots N\}$, the transition probabilities,

$$\mathbf{A} = \{a_{ijk}\} \Leftrightarrow P(S_t | S_{t-1}, I_{t-1}),$$

decompose into K separate $N \times N$ transition matrices, one for each input:

$$\mathbf{A}_k = \{a_{ij}\}_k.$$

In Sect. 16.4, we will use this structure to model the interaction between a peer tutor (input layer) and a tutee whose responses are observed (output layer). Alternatively, in Sect. 16.5, we will use the HMM in Fig. 16.1 but model the *dyadic* state rather than the tutee state as the hidden layer.

16.3 The Adaptive Peer Tutoring Assistant

We apply two DBN models to data from a development project and set of experiments by Walker and collaborators (Walker, Rummel, & Koedinger, 2009a, 2009b, 2011) at the intersection of intelligent tutoring systems and computer-supported collaborative learning (CSCL) research. The implementation objective was to provide an environment that supports high school students in learning about a particular domain through tutoring their peers. The APTA was designed to understand student tutoring actions and to provide tutors with just-in-time prompts on how to tutor more effectively. This design was informed by evidence that some amount of scaffolding support is needed to encourage productive behaviors in collaborative learners, whereas too much imposed structure may overly constrain natural behavior and/or reduce engagement (Dillenbourg, 2002; Johnson & Johnson, 1990).

Walker modified the interface to the Cognitive Tutor Algebra (Carnegie Learning) such that a peer tutor could view the progress of the tutee at work on a set of exercises. The tutee, rather than receiving feedback from the computer (as in typical use of the cognitive tutor), instead was able to request help and exchange messages with the peer tutor through a chat window. The peer tutor, meanwhile, received automated prompts/hints aimed at improving his or her tutoring. By the fourth and final study, Walker had developed both a theoretical model for productive behaviors and an algorithm for automatically assessing peer tutor chat.

Building on prior research in scaffolding peer collaboration through the use of sentence starters or classifiers, APTA encourages the peer tutor to press a button selecting “ask why,” “explain why wrong,” “hint,” or “explain next step” when providing help. Not using starters when providing help or using them when making off-topic chats is considered buggy or suboptimal behavior. The tutor may also select “other” in those cases, which counts as correctly not using starters.

Table 16.1 Frequency of automatic production rules in adaptive peer tutoring assistant data

Production rule	Count
lowLevelHelp	1162
noHelpWithNoStarters	838
noStartersWithHelp	813
helpWithHelpStarters	575
helpAfterIncorrect	212
noPromptAfterMisconception	187
noErrorFeedbackAfterMisconception	187
helpAfterRequest	144
highLevelHelp	138
startersWithNoHelp	138
noHelpAfterIncorrect	64
helpAfterCorrect	61
noHelpAfterRequestShort	43
ErrorFeedbackAfterIncorrect	30
PromptAfterIncorrect	25
helpAfterExplanation	15
noHelpAfterRequestLong	2

Positive behaviors also include providing help when needed (as well as the contrapositive, no help when not needed), providing error feedback when appropriate, and providing high-level conceptual help rather than low-level next-step help. Conditional upon prior tutee actions, the tutor chats automatically triggered 1 or more of 17 production rules, possibly several at once. These rules are compiled in Table 16.1 in decreasing order by frequency of occurrence in the complete data logs from the experiment. Not surprisingly, the label "low-level help" occurs frequently, as does "no help with no starters," which is designed to include off-task conversation.

The data set we analyze came from 124 subjects working in 62 disjoint dyads (including high school levels of Algebra 1, Geometry, and Algebra 2) for roughly 90 min (three intervals with breaks), plus instruction, preparation, and pre- and posttesting spread out over another 90 min. Tutor actions were automatically coded by the adaptive system in real time. The production rule activated by each action was determined by a combination of regular expression pattern matching and machine-learned contextual rules from earlier versions of the tutor. We use these codes in our Model 1 analysis; however, it is important to note that they have not been checked by human coders and are likely to contain inaccuracies. As we discuss, one of the advantages of the Model 1 approach is to highlight where the automatic support system may have misclassified actions. For Model 2, we consider the same sequences with human codings for both cognitive and affective characteristics and focus on the properties of the dyadic interaction instead of on the peer tutor actions.

16.4 Model 1: Direct Comparative Assistance of Tutor Utterances

Given that the tutee is under direct guidance by the tutor during the experiment, it makes sense to consider the tutee's mental state in terms of *assisted capability* rather than individual mastery. That said, if the tutee encounters an obstacle, that is, he or she makes an incorrect step in solution and corrects the mistake with the help of the tutor, we may inquire which types of tutor input provide more assistance. This is the goal of our first model, where the classes of tutor utterances are defined by the labels in Table 16.1. A data sequence of interest is thus initiated by an incorrect step and terminated by a correct step. Each dyad produces multiple sequences in the 90-min session, and we model all such sequences together across dyads. Other data, such as tutees self-correcting without assistance, are ignored.

Assistance is narrowly operationalized here as getting the tutee to the "capable" state, where a correct response is likely. This is no doubt a shortsighted definition; telling the tutee the answer is the quickest way to getting a correct output. But to start out, we accept this possibility and remain agnostic about what will be learned from the data. The graphical representation of the model is just as in the IOHMM in Fig. 16.2, except that we have relabeled the nodes in Fig. 16.3. The assistance parameter a_k for each interaction type is the probability of transitioning to a capable state from an incapable state, conditional on the interaction $I = k$. This enters as a matrix element of the transition matrix $\mathbf{A}$.

The raw XML logs from Walker's (2011) study (one for each dyad) were parsed using Python code to extract a set of sequences for analysis using the IOHMM. By design, a sequence begins at an incorrect step and ends at the next correct step (correctness of each step is evaluated and logged by the cognitive tutor). If one or more tutor production rules in Table 16.1 are triggered by tutor chats during this interval, each interaction is an observed input. When several occur simultaneously, the (arbitrary) order in which they are logged is kept; because no observed incorrect steps are possible in between, the data sequence is populated with unobserved (missing) output steps. It is, however, possible to observe more than one incorrect attempt interspersed with interactions with the tutor, as shown in Table 16.2 and,

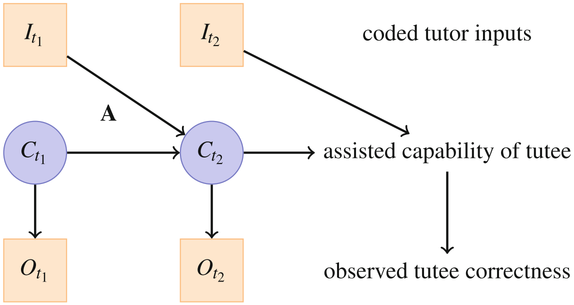


Fig. 16.3 Model 1 as an input-output hidden Markov model

Table 16.2 Sample tutor–tutee observed sequence for assistance model

Output code	Tutee's action	Tutor's action	Input code
INCORRECT	[divide rt]		
		u did that last step wrong u need to divide both sides by r + v	helpAfterIncorrect helpWithHelpStarters [5][7]
	now cuz i didnt slove for t yet [undo]		
		NO! listen to what i am saying	startersWithNoHelp [9]
INCORRECT	[divide r + t] i did now what do i have to do		
		its divide by r + t i mean r + v	lowLevelHelp, noStartersWithHelp noHelpWithNoStarters [2][8][10]
	[undo]		
CORRECT	[divide r + v]		

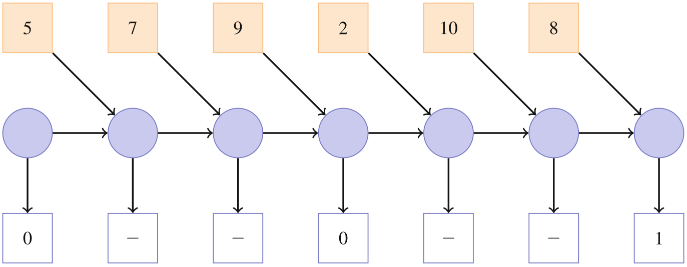


Fig. 16.4 Dynamic Bayesian network representation of a single data sequence in Table 16.2. *Square nodes* are observed, whereas *round nodes* are hidden states. For further distinction, tutor inputs are filled, whereas tutee observations (correct, incorrect, or unobserved) are hollow

graphically, in Fig. 16.4. If no interactions are observed, that is, if the student corrects a mistake without engaging with the tutor, the sequence is ignored.

Because the last seven classes in Table 16.1 occur only rarely, we ignore them in the present analysis and concentrate on the 10 most frequent labels. Parameters learned for rare events will be unreliable and may degrade the estimation of the remaining parameters. A catch-all label would be another option for handling these events—the way that rare words may be handled in natural language applications—but because some of these classes are contradictory, the catch-all event would have no interpretable value. Omitting certain events thus means that whatever assistance was provided will be redistributed among the remaining categories. This establishes that assistance is not to be interpreted in absolute terms but as a relative comparison between the included categories.

Thus defined, the data set consisted of 343 sequences, with 2169 time slices; each time slice may contain both an observed input and output. There were 10 states in the observed input layer, 2 latent states in the student-assisted capability, and 2 observable outputs. The output layer had deterministic end points—an incorrect initial state and a correct final state—with missing or incorrect states in the mid-points. All parameter learning was carried out using Murphy's (2001) Bayes Net Toolbox for Matlab, which uses a variation of the expectation-maximization (EM) algorithm. The log-likelihood manifold has local maxima, so we use multiple restarts of EM from different initial values. Using 300 restarts, we found that the 10 best runs, in terms of log-likelihood, resulted in consistent assistance values.

Some measure of the sample dependence of the assistance parameters is important, especially anticipating the wish to ask questions concerning specific subpopulations. The results of a delete-*d* jackknife subsampling procedure (Shao & Wu, 1989) are shown in Fig. 16.5, where 30 of the 62 dyads are randomly sampled without replacement 50 times. For each subsample, EM is run to convergence once using as initial values the estimates from complete data. Shown are means and standard deviations.

Labels for the interaction codes have been left off of the horizontal axis in Fig. 16.5 so as to focus the reader on the following salient features: some interactions indeed have higher assistance values according to the model, and they appear to fall roughly into two bands—values between 0.3 and 0.5 and values between 0 and 0.18. We refer to interactions in these bands as hits and misses, respectively.

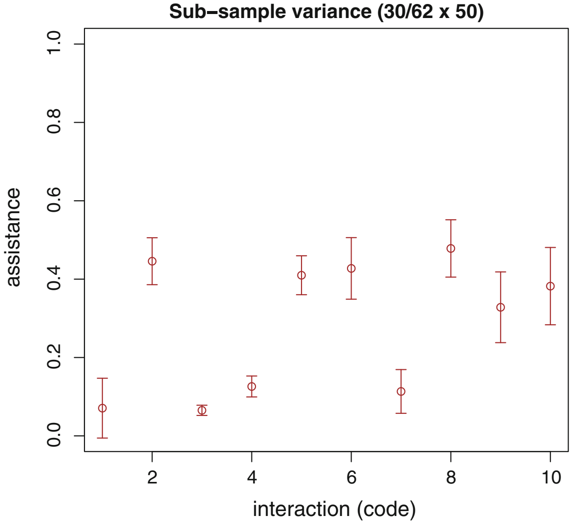


Fig. 16.5 Assistance value parameters showing subsampling variance

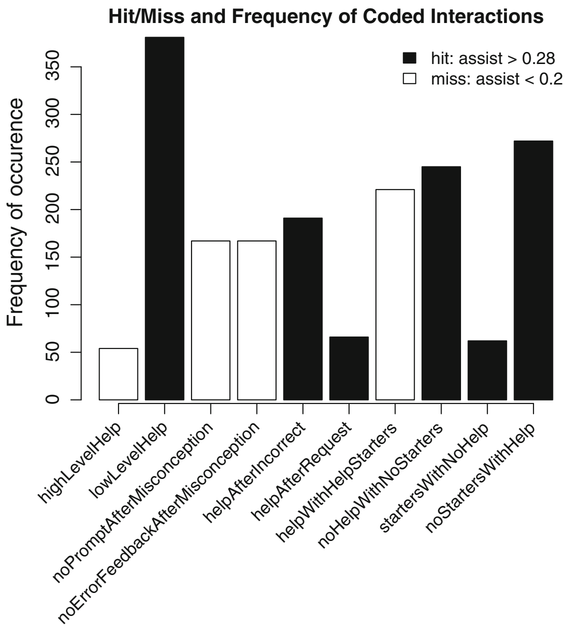


Fig. 16.6 Frequency and hit/miss classification of interactions. *Height of bars* represents frequency. *Black bars* represent interactions with high assistance (hits), whereas *white bars* represent low assistance (misses)

The bar plot in Fig. 16.6 shows both the frequency of occurrence (bar height) and the hit–miss classification (filled and empty bars, respectively) for the 10 production rules from the APTA data, now fully labeled. The order from left to right corresponds to the numbered codes in Fig. 16.5.

It is evident from Fig. 16.6 that hit–miss classification is not simply a correlate of high and low frequency of occurrence, another good sign. The classification of high-level conceptual help by the model as a miss need not clash with common sense, given that the assistance of each interaction is a measure of how quickly it aids the tutee in reaching the next correct step. Assistance must be understood here as a short-term effect, whereas abstract conceptual help may indeed play into learning more slowly. The next few interactions also appear on their face to be "correctly" classified; that is, noPromptAfterMisconception certainly sounds unhelpful in contrast to helpAfterIncorrect. However, 2 of the 10 interaction types would appear to be false positives (noHelpWithNoStarters, startersWithNoHelp) and 1 a false negative (helpWithHelpStarters).

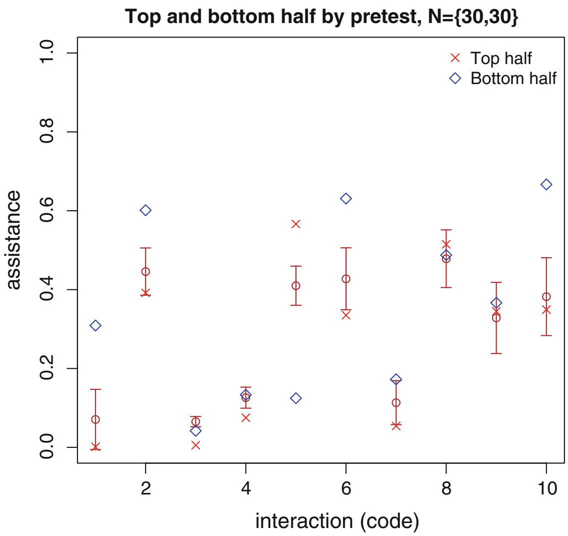
Consider first the two false positives; however agnostic one would like to be about the importance of using sentence starters, chats classified as "no help" should presumably have low assistance value. The noHelpWithNoStarters interaction occurs so frequently in the data that in 28 sequences, it is the only interaction observed. Eleven of these 28, on closer inspection, contain specific advice about how to proceed or correct a previous step ("distribute −(−y* f−yt)," "hq-mk-ks = hq-mk-ks-hq = nt-hq," "NO 16n not 16 silly willy," "sub bh not divide it"), while at least 5 contain instruction to wait or to proceed ("hold on," “solve the problem”). It appears from this sample that this interaction label suffers from inaccurate or insufficiently granular automatic coding. Inaccuracies were indeed anticipated (Walker, 2010), but now we see that the model classification has alerted us to a potential problem. The ability to detect a mismatch between expected and observed values for certain interactions may be considered a feature (as opposed to a bug) in the model. A similar story obtains for the startersWithNoHelp interaction, which does not occur by itself but does also contain both instructive chats roughly one-third of the time (“u forgot 1 lol,” “it’s actually kr,” “use subtraction”). Within these chats are also a fair number of motivational speech acts (e.g., “yes, good,” “ok I think we’re done!!”) and direct reports from the tutor that he or she has been guided by the automatic agent to mark a step taken by the tutee as incorrect (“that’s what it tells me,” “on mine it said to tell your partner that its wrong”). Thus both of the “false positives” from the model appear to be explainable in terms of shortcomings in the automatic coding, in terms of both the inaccurate classification of the instructive chats and the lack of consideration of motivational speech acts in the coding scheme.

A thornier issue surrounds the model’s assignment of low assistance value to an interaction rule labeled helpWithHelpStarters. Miscoding seems a dubious explanation, that is, that the algorithm mislabels nonhelp as help. A likelier explanation is that this interaction suffers from coincidence with other highly assistive interactions in a sequence, an example of the “explaining away” effect in Bayesian networks (and a violation of our first-order Markov assumption). Coincidence of labels occurs because the automatic coding algorithms often trigger simultaneous rules. Of the 148 sequences in which the code helpWithHelpStarters occurs, the code never occurs alone and is coincident 128 times with lowLevelHelp, a label that gets high assistance value. A few solutions exist for decoupling two interactions that are frequently coincident. One could recode the occurrence of helpWithHelpStarters differently when it co-occurs with lowLevelHelp versus when it does not,[1] or one could “clamp” the assistance value of lowLevelHelp, for example, force it to zero, and see how helpWithHelpStarters changes relative to everything else. The latter option is analogous with “controlling” for one variable, in the language of regression. It has the advantage of revealing whether other interaction codes are also significantly entangled with lowLevelHelp through coincidence. The result of pegging the assistance value of lowLevelHelp to zero is that helpWithHelpStarters indeed becomes a hit (assistance value > 0.3). Importantly the eight other interactions do not change class; the hits stay hits and the misses stay misses. Though the model was unable to decouple two of the interactions by itself, it can be coaxed into doing so, which is a partial success. Furthermore, this kind of delete-one procedure is easily automated into an analysis.

Assistance parameters using the APTA data have been shown to be computationally reliable and their clustering into hits and misses also to be fairly robust

[1] This is a special case of coding all bigrams, which is one way to recast a second-order Markov model as a first-order model.

Fig. 16.7 Differential classification of assistance parameters for tutees in the top and bottom halves of pretest scores. In addition to point estimates for each interaction (and each group), sampling variance estimates for each interaction are shown (see Fig. 16.5)



upon random subsampling of half of the dyads. Walker's experiment included pre- and posttests of the participants, which invites at least one last analysis. If some interaction types are more assistive at higher or lower levels of tutee ability, this might be discoverable by restricting the data set to the top and bottom 30 dyads by tutee pretest score, as shown in Fig. 16.7.

Assistance values of the interactions for the top prescoring tutee dyads (red Xs) appear to be consistent with random subsamples (brown error bars from summary statistics shown in Fig. 16.5) as well as with the whole data set. For the bottom-scoring tutee dyads (blue diamonds), 2 of the 10 interactions appear to change class: highLevelHelp changes from a miss to a hit and helpAfterIncorrect from a hit to a miss. The first of these seems to suggest that conceptual help is more assistive for the lowest prepared tutee, at least in the short term. A possible explanation is that better prepared tutees (higher pretest scores) are aware of the procedures but get stuck predominantly on their implementation. The flipping of helpAfterIncorrect also makes sense. A tutee who is prone to mistakes on almost every step will hardly benefit extra from being made aware of them, whereas a somewhat capable tutee can constructively incorporate such feedback. The frequency of observed events in the logs was shown in Table 16.1, but we point out here that in differentiating between the bottom- and top-scoring tutees, estimates of assistance of highLevelHelp were based on 26 observations in each group, which is a modest amount. Estimates of the assistance for helpAfterIncorrect were based on 82 and 107 observations in the bottom and top group, respectively.

The principal findings from Model 1 using the APTA data were as follows: (a) The inference of assistance parameters for the interactions was computationally reliable and stable under resampling of the data; (b) interactions appeared to cluster naturally into two classes of high and low assistance, that is, hits and misses;

(c) upon comparison of the interaction classification with the interaction codes, two apparent false positives were explained by inaccuracies or other insufficient granularity of the coding; (d) an apparent false negative was attributed to strong coupling by coincidence with another interaction, but a ready solution to this confound did bring the model estimates in line with expectations; and (e) comparing the relative assistance values for tutees in the bottom and top brackets by pretest scores, the hits and misses were unchanged for the top bracket, whereas 2 of the 10 flipped in the bottom bracket, a possible indication of differential cognitive benefits.

Inaccuracies in the automatic coding of the interactions is one deficiency in the first analysis. Another is that in the input–output model, where inputs corresponded to tutor actions and outputs to tutee correctness, there were typically many more inputs than outputs. To line up the sequences by time slice, the output sequence was padded with "unobserved" values, that is, missing data. But neither are the data missing at random nor is "missingness" explainable by any parameter in the model. For example, one might consider augmenting the observed output of the tutee to include not only correct–incorrect but also a category for "declined to answer." This would make sense if turn taking were deliberately structured in the tutor–tutee interaction, which was not the case. Because tutor chats triggered multiple codes at the same time, a missing attempt could be artifactual. Both of these issues are addressed in the application of Model 2, described in the next section.

16.5 Model 2: A Discriminative Hidden Markov Model to Predict Learning Gain

The second model we apply to the APTA data set is designed to characterize the dyadic patterns rather than the flow of assistance from the tutor to the tutee. While the latent variable in Model 1 was tutee mastery, the latent variable is now state of dyad, which is not prescribed by theory but rather discovered in an exploratory fashion. The interpretation of this hidden state must come from examining the various emission probabilities associated with it. The cardinality of the state will be determined empirically, as we shall show.

Because all observations, tutor or tutee, are emissions from a hidden dyadic state, the data sequences do not need to be padded with unobserved values. The affective coding used here is a repurposed subset of codes from a study on affect (Ogan, Finkelstein, & Walker, 2012) that included both tutor and tutee chats. However, because the tutee chats were not coded for cognitive labels, all tutee chats have been aggregated under the generic label "chat." The frequency counts of all of the codes used are shown in Table 16.3.

The raw XML logs from APTA (one for each dyad) were parsed using Python code to extract events of interest, and data were merged with cognitive and affective coded chat tables using the unique chat strings. Note that although cognitive codes in Table 16.3 were mutually exclusive, affective codes could occur in combination

Table 16.3 Frequency of action codes in adaptive peer tutoring assistant data

Actor	Code	Count
Tutee	Incorrect	1380
	Correct	1690
	Undo	2544
	Chat	2333
Tutor (cognitive)	wrongstep.feedback	132
	elaborated.explanation	69
	elaborated.hint	181
	unelaborated.explanation	1189
	unelaborated.hint	168
Tutor (affective)	Laughter	150
	Positivity	456
	Impoliteness	219
	Rudeness	160
	nocode (off-topic)	712

with cognitive or other affective codes. Because our HMM approach requires a unique observable, we included the 13 most frequently occurring combinations of codes in the data for analysis. Thus “Impoliteness” is included in addition to “Impoliteness and Rudeness,” which was more frequent than “Rudeness” alone. Including the four tutee observable states (incorrect, correct, undo, chat), there were thus 17 possible output states.

The conversions of data into sequences for the HMM estimation is illustrated in Table 16.4 and Fig. 16.8. There were 10,800 observed outputs in the data set for all of the dyads, each of whose sequences ranged in length from 93 to 327 actions.

To build a classifier, the dyads were first sorted into high- and low-gain groups based on the normalized gain of the tutee from pre- to posttest, $g = (S_{\text{post}} - S_{\text{pre}})/(1 - S_{\text{pre}})$. High gain correspondoed to $g > 10\%$ (20 dyads, $\bar{g} = 28\%$), low gain to $g \leq 0\%$ (20 dyads, $\bar{g} = -3\%$). As the high-gain and low-gain sample groups are the same size, the baseline comparison for a classifier is indeed 50%.

We used leave-one-out cross-validation, that is, leaving out one dyad at a time. A separate HMM was learned for all of the dyads in the high-gain group and for the low-gain group. The left-out sequence was then classified by the likelihood of the data determined from each of the models. A schematic is shown in Fig. 16.9. The cardinality of the hidden state was varied from 2 to 10. In addition, full-data models for each hidden state cardinality were trained so that information criteria, AIC and BIC, could be computed to evaluate model fit.

Model parameters were learned using the Bayes Net Toolbox for Matlab (Murphy, 2001). Because the log-likelihood surface for a HMM is nonconcave, parameter estimation is susceptible to local maxima. In practice, this means that estimation is restarted multiple times (we used 50 restarts) from random parameter values and allowed to run for several cycles of EM (we used 20). Parameters from the run with the highest final log-likelihood were kept. We used fewer restarts than

Table 16.4 Sample tutor–tutee observed sequence for learning gain model

Tutee codes	Tutee's action	Tutor's action	Tutor code (plus affect)
INCORRECT	[divide rt]		
		u did that last step wrong u need to divide both sides by r + v	wrongstep.feedback unelaborated.explanation
CHAT UNDO	now cuz i didnt slove for t yet [undo]		
		NO! listen to what i am saying	Impoliteness, Rudeness
INCORRECT CHAT	[divide r + t] i did now what do i have to do		
		its divide by r + t i mean r + v	unelaborated.explanation unelaborated.explanation
UNDO CORRECT	[undo] [divide r + v]		

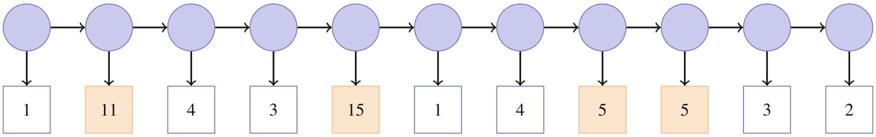
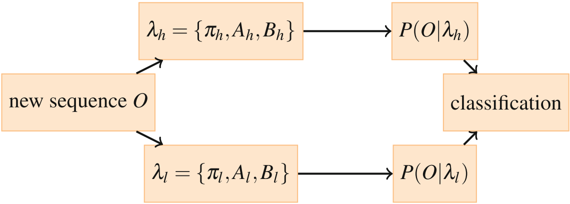


Fig. 16.8 Dynamic Bayesian network representation of the data sequence excerpt in Table 16.4. *Square nodes* are observed, whereas *round nodes* are hidden states. For further visual distinction, tutor inputs are filled, whereas tutee actions are hollow. The full sequence for this dyad is considerably longer

Fig. 16.9 Schematic of hidden Markov model classifier


$\lambda_h = \{\pi_h, A_h, B_h\}$
$P(O|\lambda_h)$
new sequence O
classification
$\lambda_l = \{\pi_l, A_l, B_l\}$
$P(O|\lambda_l)$

we did for Model 1, because the leave-one-out procedure meant repeating each estimation 19 times. Results are shown in Fig. 16.10.

The best cross-validated performance (78% accuracy) was observed for a model with a hidden state cardinality of 8, which is consistent with the AIC measure. BIC appears to overpenalize the additional parameters of the more complex model.

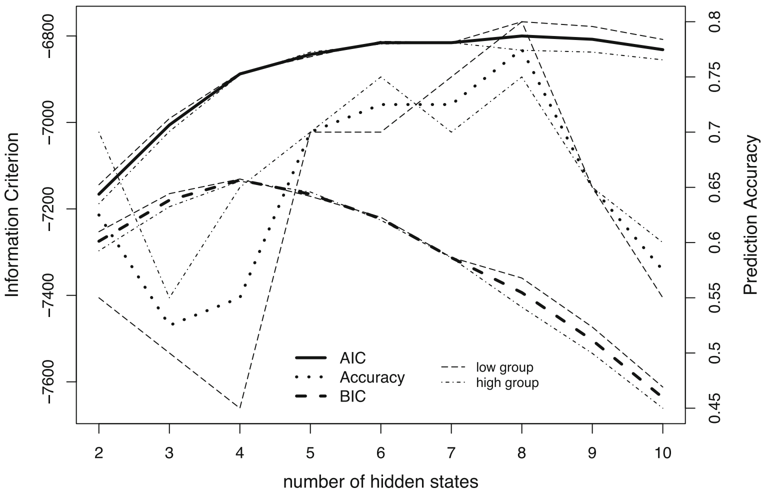


Fig. 16.10 Model selection using information criteria and cross-validated accuracy. Wide types distinguish, from top to bottom, AIC, Rand accuracy, and BIC. Thin line types distinguish low and high subgroups for each criterion

For comparison, we build two logistic regression classifiers based on counts of the codes in Table 16.3 for each dyad in the high and low groups. The naive logistic regression classifier has no model-selection phase but simply uses all of the available data during each leave-one-out iteration. Alternately, we first perform stepwise forward-backward model selection (stepAIC in R) on the full data set and then evaluate the resulting model error using leave-one-out cross-validation. The best model included only the following predictors of normalized gain: Incorrect, Undo, elaborated.hint, elaborated.explanation, unelaborated.explanation, Laughter, and Positivity. Cross-tabulated results for these classifiers and for the best HMM are shown in Table 16.5. Also shown are Goodman and Kruskal's (1954) λ measure of proportional reduction of error, with confidence intervals.

As seen from Table 16.5, the sample sizes in this analysis are too small for the differences between models to be statistically significant, although the HMM is the only classifier that is significantly better than chance ($\lambda = 0$). With these table proportions, at a 95% confidence level, we would have needed roughly 10 times the number of dyads to distinguish between the best logistic model and the HMM. Conversely, at the present sample size, the HMM classifier would have needed to correctly classify at least 19 out of 20 dyads in each class, which was not a realistic expectation. Therefore, from the point of view of beating the logistic classifier, the study was underpowered. It may indeed be the case that the dynamic model is a superior classifier, taking advantage of recurring patterns within the output sequence that are washed away if the states are simply aggregated. That said, a logistic-regression classifier could also be built on bigram (or trigram, etc.) frequencies from the data set.

Table 16.5 Cross-tabulated performance

Classifier	Predicted label	Actual		λ (95% conf)
		High	Low	
Naive logistic	high	11	9	0.10 (0.00, 0.52)
	low	9	11	
Best logistic	high	12	5	0.35 (0.00, 0.73)
	low	8	15	
Hidden Markov model	high	16	5	0.55 (0.26, 0.84)
	low	4	15	

Interpreting the hidden states of the learned HMM—in particular, what it is that makes the high-gain group successful—is the more interesting work. We take a few steps in that direction here, although it is by no means a complete explanation. For illustrative purposes, we include the 8 × 8 transition matrix (i.e., between hidden states) for the high-gain group in Table 16.6.

Notably from Table 16.6, State 5 appears to be a very stable state; that is, there is an 85% chance of remaining in State 5 in the Markov chain. This same state has a 89% emission probability of a tutee undo. Thus it is reasonable to identify State 5 as the undoing state; it is stable because one undo operation is very often followed by another. Along the diagonal, State 8 is also fairly stable. It is one of two hidden states (along with State 1) strongly connected to correct responses. It should not be surprising that a series of correct steps is also a common pattern.

A more subtle feature in Table 16.6 is a jumping process or oscillation between States 4 and 7. State 4 has a 53% chance of transitioning to State 7, which then has a 68% chance of transitioning back to State 4. Examination of the 8 × 17 observation matrix suggests that this is off-topic chat between the tutor and the tutee (hence requiring two states). Observations of this kind support the use of bigram analysis as a prescreening phase, as was done in Boyer et al. (2011). In addition to an emission probability of 40% of off-topic chat, State 4 also has a 19% probability of positive chat, suggesting that perhaps this oscillation represents a rapport-building exchange.

Two of the tutor–tutee "modes," a stable (tutee) undo state and off-topic chat, can be found in the low-gain transition matrix, though the nominal state numbers are of course different. Thus, although they are interpretable, the mere existence of these modes does not help to distinguish high-gain behavior from low-gain behavior. The series of correct responses pattern is notably absent from the low-gain matrix, where, in contrast, a cycle appears between unelaborated explanations and correct responses (suggestive of feeding the answers to the tutee). This cycle is congruent with the finding in Sect. 16.4 that low-level help had a high assistance score. The role of high-level help (elaborated hints and explanations), which had a high assistance score for low-pretest students, is less clear in this model, perhaps because of its infrequent occurrence.

Whether relative frequency of these modes and/or other interpretable modes can be used to understand productive tutoring sessions—and ultimately provide automatic feedback in peer learning contexts—is the subject of continued investigation.

Table 16.6 Hidden state transition matrix for the high-gain group

	1	2	3	4	5	6	7	8
1	0.3923	0.0076	0.1578	0.0567	0.0001	0.0015	0.0878	0.2962
2	0.0341	0.0639	0.1428	0.1860	0.0749	0.1429	0.3421	0.0132
3	0.0037	0.0722	0.4405	0.0054	0.1067	0.3175	0.0523	0.0017
4	0.1018	0.2128	0.0586	0.0777	0.0011	0.0033	**0.5274**	0.0172
5	0.0134	0.0387	0.0125	0.0004	**0.8482**	0.0386	0.0045	0.0436
6	0.0253	0.0548	0.5212	0.0039	0.0553	0.0062	0.0209	0.3124
7	0.0378	0.1101	0.0512	**0.6825**	0.0027	0.0408	0.0188	0.0561
8	0.2607	0.0025	0.0871	0.0079	0.0001	0.0023	0.0614	**0.5778**

Note Some larger values are in boldface for emphasis

16.6 Conclusions and Future Work

We have presented two different ways of modeling sequential data from collaborative interactions using DBNs. In Model 1, coded tutor chats were modeled as inputs, whereas tutee correctness was the only type of modeled output. The purpose of this model was to infer comparative assistance values of different tutor utterance classes on the tutee's assisted capability (binary hidden state) to correct a mistake. Although the model was somewhat successful, imprecise real-time automated codes and missing data limited the model's usefulness.

Second, we modeled the dyadic state ($|S| = 8$) rather than the tutee state as a hidden layer with a discriminative HMM approach to classify high and low learning gains. This time, there were 17 observed action categories, comprising both cognitive and affective labels that were human coded. Classification accuracy of the HMM (78%) exceeded the best static logistic regression model (68%), although sample sizes were too small for this difference to be significant. We acknowledge that further investigation is needed to make stronger claims about the value of dynamic information in modeling peer interactions.

Although some features of the learned models are straightforward to interpret, much work remains to understand the differences between high- and low-gain groups in such a way that actionable interventions are possible. Exploring interactions between cognitive and affective factors is also an interesting direction. Revisiting theories of learning from peer tutoring, and leveraging these theories to iterate on the codes used to characterize tutor and tutee dialogues, may yield more interpretable results with clearer connections across the two approaches. We consider our contribution to be a proof-of-concept of how HMM approaches can be used to extract patterns of interest in collaborative interactions, such as may be used to detect effective and ineffective collaborative interactions and, where appropriate, trigger adaptive support.

References

Almond, R. G. (2007). *An illustration of the use of Markov decision processes to represent student growth (learning)* (ETS Research and Development Report No. RR-07-40). Princeton, NJ, USA: ETS

Baghaei, N., Mitrovic, A., & Irwin, W. (2007). Supporting collaborative learning and problem-solving in a constraint-based CSCL environment for UML class diagrams. *International Journal of Computer-Supported Collaborative Learning, 2*(2–3), 159–190. doi:10.1007/s11412-007-9018-0

Baum, L., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics, 37*(6), 1554–1563.

Bengio, Y., & Frasconi, P. (1995). An input output HMM architecture. In *Advances in neural information processing systems* (pp. 427–434). Cambridge, MA, USA: MIT Press.

Blackwell, D., & Koopmans, L. (1957). On the identifiability problem for functions of finite Markov chains. *The Annals of Mathematical Statistics, 28*(4), 1011–1015.

Boyer, K. E., Phillips, R., Ingram, A., & Ha, E. (2011). Investigating the relationship between dialogue structure and tutoring effectiveness: A hidden Markov modeling approach. *International Journal of Artificial Intelligence in Education, 21,* 65–81.

Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. In P. A. Kirshner (Ed.), *Three worlds of CSCL: Can we support CSCL?* (pp. 61–91). Heerlen, Netherlands: Open Universiteit Nederland.

Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence, 15*(1), 9–42.

Gilbert, E. (1959). On the identifiability problem for functions of finite Markov chains. *The Annals of Mathematical Statistics, 30*(3), 688–697.

Goodman, L., & Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*(268), 732–764.

Johnson, D. W., & Johnson, R. T. (1990). Cooperative learning and achievement. In S. Sharan (Ed.), *Cooperative learning: Theory and research* (pp. 23–37). New York, NY, USA: Praeger.

Kumar, R., Rosé, C., Wang, Y., Joshi, M., & Robinson, A. (2007). Tutorial dialogue as adaptive collaborative learning support. In *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building technology rich learning contexts that work* (pp. 383–390). Amsterdam, The Netherlands: IOS Press.

Magnisalis, I., Demetriadis, S., & Karakostas, A. (2011). Adaptive and intelligent systems for collaborative learning support: A review of the field. *IEEE Transactions in Learning Technologies, 4*(1), 5–20. doi:10.1109/tlt.2011.2

Murphy, K. P. (2001). The Bayes net toolbox for Matlab. *Computing Science and Statistics, 33*(2). Retrieved from https://code.google.com/p/bnt/

Nag, R., Wong, K., & Fallside, F. (1986). Script recognition using hidden Markov models. In *ICASSP '86: IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 11, pp. 2071–2074). New York, NY, USA: Institute of Electrical and Electronics Engineers. doi:10.1109/ICASSP.1986.1168951

Ogan, A., Finkelstein, S., & Walker, E. (2012). *Rudeness and rapport: Insults and learning gains in peer tutoring*. Paper presented at the 11th International Intelligent Tutoring Systems Conference, Chania, Greece.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education, 14,* 1–33.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In *Handbook of research on mathematics teaching and learning* (pp. 334–370). New York, NY, USA: Macmillan.

Shao, J., & Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics, 17*(3), 1176–1197. doi:10.1214/aos/1176347263

Soller, A., & Stevens, R. (2007). *Applications of stochastic analyses for collaborative learning and cognitive assessment*. Alexandria, Virginia, Egypt: Technical report.

Starner, T., & Pentland, A. S. (1997). Real-time american sign language recognition hidden Markov models from video using. In M. Shah & R. Jain (Eds.), *Motion-based recognition* (pp. 227–243). Amsterdam, The Netherlands: Springer.

Walker, E. (2010). *Automated adaptive support for peer tutoring* (Dissertation, Carnegie Mellon University).

Walker, E., Rummel, N., & Koedinger, K. R. (2009a). CTRL: A research framework for providing adaptive collaborative learning support. *User Modeling and User-Adapted Interaction, 19*(5), 387–431. doi:10.1007/s11257-009-9069-1

Walker, E., Rummel, N., & Koedinger, K. R. (2009b). Modeling helping behavior in an intelligent tutor for peer tutoring. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graessar (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 341–349). Amsterdam, The Netherlands: IOS Press.

Walker, E., Rummel, N., & Koedinger, K. R. (2011). Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. *International Journal of Computer-Supported Collaborative Learning, 6*(2), 279–306. doi:10.1007/s11412-011-9111-2

Walker, E., Rummel, N., & Koedinger, K. R. (2014). Adaptive intelligent support to improve peer tutoring in algebra. *International Journal of Artificial Intelligence in Education, 24*(1), 33–61. doi:10.1007/s40593-013-0001-9

Chapter 17
Representing Self-organization and Nonstationarities in Dyadic Interaction Processes Using Dynamic Systems Modeling Techniques

Sy-Miin Chow, Lu Ou, Jeffrey F. Cohn, and Daniel S. Messinger

Abstract Dynamic systems modeling techniques provide a convenient platform for representing multidimensional and multidirectional change processes over time. Central to dynamic systems models is the notion that a system may show emergent properties that allow the system to self-organize into qualitatively distinct states through temporal fluctuations in selected key parameters of interest. Using computer vision-based measurement of smiling in one infant-mother dyad's interactions during a face-to-face interaction, we illustrate the use of generalized additive modeling techniques to fit multivariate dynamic systems models with self-organizing, time-moderated dynamic parameters. We found evidence for systematic over-time changes in the infant → mother cross-regression effect, which provided a glimpse into how the dyad self-organized into distinct states over the course of the interaction, including periods where the mother's positivity was reinforced and strengthened by the infant's positivity, as well as periods where the mother's positivity was inversely related to the infant's past positivity levels.

Keywords Dynamic systems · Time-varying coefficients · Self-organization · Dyadic interaction · Generalized additive model · Face-to-face/still-face · Nonstationarity

S.-M. Chow (✉) · L. Ou
Pennsylvania State University, State College, USA
e-mail: symiin@psu.edu

L. Ou
e-mail: lzo114@psu.edu

J.F. Cohn
University of Pittsburgh, Pittsburgh, USA
e-mail: jeffcohn@pitt.edu

D.S. Messinger
University of Miami, Coral Gables, USA
e-mail: dmessinger@miami.edu

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_17

17.1 Introduction

Modeling how group members act, react, and interact with each other is a challenging and inherently high-dimensional problem (Mislevy et al., 2014). Recent years have seen a rapid growth of interest in conceptualizing group collaborations as dynamic processes that evolve over time. For instance, Soller and Stevens (2007) used hidden Markov models to represent online knowledge sharing as a sequence of transitions between effective knowledge-sharing episodes and knowledge-sharing breakdowns. In lower-dimensional settings, Halpin and colleagues (see Chap. 17, this volume; see also Halpin & Deboeck, 2013) have used the Hawkes process to represent dyadic interaction as a dynamic process wherein the actions of each individual affect the dyad's probability of further actions.

One challenging but critical aspect to address in modeling group dynamics concerns whether and how researchers can effectively represent real-time changes in the relations among team members as situational changes unfold (e.g., Hao et al., Chap. 17, this volume, for an application involving the Tetralogue chat during which communications and collaborations among team members unfold). For instance, at the start of a collaborative game, members of the team may discuss and exchange strategies with a few close neighbors via reciprocal but amicable exchanges. As the time pressure kicks in, the exchanges may become more intense or even escalate into disagreements or arguments. Particular members of the team may begin to emerge as team leaders and exert much stronger influence on the team than they are influenced by their team members. In the terminology of dynamic systems, the team is said to have *self-organized* into qualitatively distinct states through changes in selected key parameters or variables of interest (e.g., time pressure in this particular example; Barton, 1994; Haken, 1977/83; Kelso, 1995; Smith & Thelen, 1993; Thelen, 1989). In this chapter, we illustrate instances of self-organizing properties through the special case of parent-child interaction—a dyadic process. Even though we restrict ourselves here to the lower-dimensional problem of modeling a dyadic process, tools for exploring the self-organizing dynamics of two individuals provide a fundamental building block for evaluations involving larger teams.

Parent-child influence encompasses both infant-to-parent (parental responsivity) and parent-to-infant (infant responsivity) influence (Brazelton, Koslowski, & Main, 1974). Moderate to high parental responsivity has been found to be associated with the development of secure infant attachment to the parent (Isabella & Belsky, 1991; Jaffe et al., 2001), conscience-based rule-following in the child (Kochanska, Forman, & Coy, 1999), the infant's understanding of developing emotional expressions (Stern, 1985; Tronick, 1989), as well as linguistic and cognitive development (Feldman & Greenbaum, 1997; Feldman, Greenbaum, Yirmiya, & Mayes, 1996; Landry, Smith, Miller-Loncar, & Swank, 1997). The emergence of infants' ability to adapt to changes in their parents' behavioral and emotional patterns has also been regarded as a developmental milestone (Ainsworth, Blehar, Waters, & Wall, 1978; Brazelton et al., 1974; Tronick & Gianino, 1986). Previous studies have reported evidence of time-varying concurrent infant-parent synchrony on a second-by-second basis (i.e., within episodes of a face-to-face/still-face procedure; Chow, Haltigan, & Messinger, 2010), and shown that changes in infant social engagement within the still face episode—a brief episode during which the parent ceases interaction and maintains a neutral expression toward the infant—were associated with infant attachment security and internalizing problems (Ekas, Haltigan, & Messinger, 2013).

No study has, to date, simultaneously investigated the over-time constancy of parental and infant influence on each other in real-time interaction. Such within- and between-person heterogeneities[1] in change constitute a source of nonstationarities[2] that, if left unmodeled, can obscure and distort our understanding of the dynamics of the system. The lack of readily accessible tools for diagnosing and evaluating the nature of such time-varying infant-parent relations is one reason for the scarcity of modeling work along these lines (Chow, Hamaker, & Allaire, 2009; Chow, Zu, Shifren, & Zhang, 2011; De Jong & Penzer, 1998). The lack of concrete empirical evidence for changes in bidirectional influence, in turn, calls for the need to first explore the functional forms of such changes before confirmatory approaches are used (Chow et al., 2011; Molenaar, 1994).

Much of the pioneering and seminal work on models with time- and state-dependent parameters originated from and was further popularized in the econometric, statistical, and physiological modeling literature. Examples of such models include, but are not limited to, univariate time-varying parameter models such as the local linear trend model (Durbin & Koopman, 2001), the time-varying autoregressive moving average (ARMA) model (Tarvainen, Georgiadis, Ranta-aho, & Karjalainen, 2006; Weiss, 1985), the stochastic regression model (Pagan, 1980) and dynamic factor analysis models with time-varying factor loadings or dynamic parameters (Chow et al., 2011; Del Negro & Otrok, 2008; Molenaar, 1994; Stock & Watson, 2008). In most cases, selected parameters from the model are allowed to vary contingent on time or other variables in the system, and a function deemed flexible enough to capture variations in the parameters is then incorporated into the original dynamic model.

In this chapter, we use the flexible smoothing and estimation routines for fitting generalized additive mixed models (GAMM), available as part of the R package MGCV (Heywood, Cornelius, & Carver, 2006), to diagnose, evaluate, and represent self-organizing dynamics. The general GAMM framework, which extends the generalized linear model (McCullagh & Nelder, 1989) and generalized additive model (Hastie & Tibshirani, 1990; James, 2002), postulates that person i's response

[1] Heterogeneities may stem from (a) between-person differences in the population due, for example, to the presence of subgroups/subpopulations or other individual difference characteristics; and (b) within-person variations in change characteristics during portions of the individual's repeated assessments.

[2] Strict stationarity refers to the property that the probability distribution of a stochastic process is assumed to be constant over time, whereas weak stationarity only requires the mean and variance of a probability distribution to be time invariant (Chatfield, 2004).

variable, y_i (where i = 1, …, n, with n indexing the total number of subjects), may be distributed as a member of the exponential family (e.g., normal, Poisson, gamma, multinomial, etc.; for further examples see Chap. 13, Cohen, Cohen, West, & Aiken, 2003). The mean of y_i, $\mu_i \equiv \mathrm{E}(y_i)$, is linked to a semiparametric predictor, η_i, expressed as:

$$\begin{aligned}\eta_i = \boldsymbol{X}_i\boldsymbol{\beta} + \sum_{k=1}^{K} f_{1,k}(x_{1,ki}) + \sum_{o=1}^{O}\sum_{r=1}^{R} f_{2,o}(x_{2,ri})x^*_{2,oi} \\ + \sum_{q=1}^{Q}\sum_{s=1}^{S} f_{3,s,q}(x_{3,qi}, x^*_{3,si}) + \boldsymbol{Z_i b_i},\end{aligned} \tag{17.1}$$

via $\eta_i = g(\mu_i)$, where g is a link function that maps the mean of y_i to η_i and $g^{-1}(\eta_i)$ is the reverse transformation that converts η_i into μ_i. The first and last terms constitute the usual parametric components in standard linear mixed effects models; the second, third, and fourth terms are nonparametric components wherein the effects of a series of covariates on the mean of the dependent response variable are of unknown functional forms. Specifically, $\boldsymbol{X}_i$ is the design matrix that contains person i's fixed effects components, and $\boldsymbol{\beta}$ is the corresponding vector of fixed effects parameters; $\boldsymbol{Z_i}$ is the random effects design matrix for person i, and $\boldsymbol{b}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi}_\mathbf{b})$ is a vector of random effects assumed to be multivariate normally distributed with zero means, and covariance matrix, $\boldsymbol{\psi}_\mathbf{b}$. $x_{1,ki}$ $(k = 1, 2, \ldots, K)$, $x_{2,ri}$ $(r = 1, \ldots, R)$, $x^*_{2,o,i}$ $(o = 1, \ldots, O)$, $x_{3,qi}$ $(q = 1, 2, \ldots, Q)$, and $x^*_{3,si}$ $(s = 1, 2, \ldots, S)$ are person-specific covariates. The term $f_{1,k}$ is the smooth function of the kth covariate, $x_{1,ki}$ $(k = 1, \ldots, K)$. Expanding the example on team dynamics noted earlier, researchers may wish to evaluate a possible nonlinear time trend in each team member's level of performance by including time as one of the K covariates. Alternatively, covariates such as stress may have a nonlinear effect on each member's level of performance (Henderson, Snyder, Gupta, & Banich, 2012) and may be included as another covariate in this particular smooth term.

The third term, $f_{2,o}$ (o = 1, …, O), represents O smooth functions that allow the covariates $x_{2,ri}$ to have smoothly varying interaction effect with another (unsmoothed) covariate $x^*_{2,oi}$ (Hastie & Tibshirani, 1993; Ibrahim, Leelahanon, & Li, 2005). In other words, this set of smooth functions allows the effects of the covariates $x_{2,ri}$ $(r = 1, \ldots, R)$—for instance, stress—to vary smoothly at each value of person i's oth covariate, $x^*_{2,oi}$, where the latter is typically a discrete-valued covariate such as sex, discrete time, or geographical region[3]. Thus, this smooth term

[3] GAMM provides a collection of procedures for approximating these functions and the resultant curves using different smoothers. $f_{1,k}(.) - f_{3,s,q}$ are typically referred to as smooth functions, and the curves or lines produced by these functions are denoted as smooths (Hastie & Tibshirani, 1990; McKeown & Sneddon, 2014). Note that the first subscript in $x_{1,ki} - x^*_{3,si}$ is used to distinguish the kind of smoothing function with which a specific covariate is associated, and the second subscript distinguishes among the covariates that are subjected to that particular kind of

provides an easy way of testing whether and how the covariates in $x^*_{2,oi}$ may moderate the possibly nonlinear effects of the covariates in $x^*_{2,oi}$. As an example, researchers may use this smooth term to find evidence for gender or over-time differences in the possible nonlinear effects of stress.

The terms $f_{3,s,q}$ ($s = 1, \ldots, S$; $q = 1, \ldots, Q$) are SQ smooths of tensor products used to approximate the unknown but jointly nonlinear effects of pairs of covariates on $\boldsymbol{\eta}_i$ (e.g., the interaction between $x_{3,qi}$ and $x^*_{3,si}$). Building on the example on stress and individual performance level, the roles of stress and time on member performance may show nonlinear dependencies on each other, such that a team member's level of performance may be relatively constant over time at low levels of stress but show increased inconsistencies—alternating between weak and strong performance—at high levels of stress.

While Eq. 17.1 is univariate (i.e., featuring only a single dependent variable), multivariate extensions of the model shown in (17.1) may be implemented using the dummy indicator approach frequently utilized in bivariate mixed effects models (MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997), thus allowing the GAMM modeling framework to be used to model multivariate dependencies among multiple members of a team. Due to the generality of the GAMM framework in handling any of the densities from the exponential family, the approach illustrated here, by extension, can also be used for model exploration with data that are not normally distributed (e.g., categorical and count data) provided they conform to the exponential family. Here, we will limit ourselves to the special case of continuous data from a single dyad within one particular face-to-face interaction episode. Thus, normal distributions are assumed for the responses (i.e., $\eta_i = \mu_i$) and no random affects are included through $\boldsymbol{b}_i$.

17.2 Method

The data utilized for modeling in this article were previously published elsewhere to demonstrate the utility of automated measures of mother-infant facial expression during face-to-face interaction, using techniques that are distinct from the GAMM approach undertaken here (Messinger, Mahoor, Chow, & Cohn, 2009). The data used for our modeling purposes were facial movement measures from one dyad consisting of a 6-month-old infant and his mother engaging in face-to-face interaction with unobstructed, close to full frontal views of each partner's face over 43 s. The data were acquired on a frame-by-frame basis (30 frames per second, with a

(Footnote 3 continued)

smoothing function. The superscript * in $x^*_{2,o,i}$ and $x^*_{3,si}$ is used to distinguish between the two sets of covariates that appear in the second and third types of smoothing functions. For instance, $x_{2,ri}$ denotes the rth covariate that is subjected to smoothing in $f_{2,o}$ whereas $x^*_{2,o,i}$ denotes the oth (unsmoothed) covariate that moderates the effect of $f_{2,o}\left(x_{2,ri}\right)$ on η_i.

total of 1292 frames) and aggregated into 0.1-s intervals (i.e., every 3 frames), yielding a total of 430 time points for model-fitting purposes. We used smile strength and eye constriction measures, obtained using computer vision software, CMU/Pitt's Automated Facial Image Analysis (AFA4), which produced the Facial Action Coding System (Ekman & Friesen, 1978) intensity measures. Mouth opening was measured as the vertical distance between the upper and lower lips. All measures were normalized and expressed as Z scores. Based on results from previous analysis (Messinger et al., 2009), we computed composite positivity scores for the infant and the mother (denoted below as *Infant* and *Mom*, respectively) using measures that were found to show strong convergent associations with continuous rating of positive emotion. These measures included levels of smile strength, mouth opening, and eye constriction for the infant, and smile strength and eye constriction for the mother. Plots of the automated measures used to create the composite positivity scores as well as the dyad's composite positivity scores over time are shown in Fig. 17.1a–c.

We began by exploring a linear parametric model in the form of a vector autoregressive model of order 2 (i.e., a VAR(17.2) model), denoted as Model 1, as:
Model 1

$$\begin{aligned}\mu_{Mom,t} &= \beta_{M1} Tickle_t + f_{M1}(Time_t) + \beta_{M2} Mom_{t-1} + \beta_{M3} Mom_{t-2} + \beta_{M4} Infant_{t-1}\\ \mu_{Infant,t} &= \beta_{I1} Tickle_t + f_{I1}(Time_t) + \beta_{I2} Infant_{t-1} + \beta_{I3} Infant_{t-2} + \beta_{I4} Mom_{t-1}\end{aligned} \tag{17.2}$$

where $\mu_{Mom,t}$ and $\mu_{Infant,t}$ denote, respectively, the mother's and infant's expected level of positivity at time t, $Time_t$ is an index of time, and $Tickle_t$ is a binary indicator of whether the mother was tickling the infant at time t (0 = no, 1 = yes). β_{M2} and β_{I2} represent the linear lag-1 autoregression effect of each dyad member's previous positivity level at time $t-1$ on his/her own positivity level at time t, whereas β_{M3} and β_{I3} are the corresponding linear lag-2 autoregression effects; in a similar vein, β_{M4} and β_{I4} represent the linear lag-1 cross-regression effects of the other dyad member's positivity level on a dyad member's current positivity level. These cross-regression effects are constrained to be time-invariant in this model and serve as general indices of the extent to which one dyad member's positivity is coupled to the other dyad member's positivity. Thus, if the mother's (or infant's) positivity is coupled to, or is affected by the infant's (mother's) positivity at a previous time point, namely, β_{M4} (or β_{I4}) is statistically different from zero, the mother (infant) is said to be generally—or on average—responsive to fluctuations in the infant's (mother's) positivity over the course of the entire interaction.

A smooth of the time trend manifested by the two dyad members is included for the mother and infant, respectively, through $f_{M1}(Time_t)$ and $f_{I1}(Time_t)$. These two smooth terms allowed us to capture nonparametric but relatively slow fluctuations in each dyad member's positivity level that unfolded independently of the

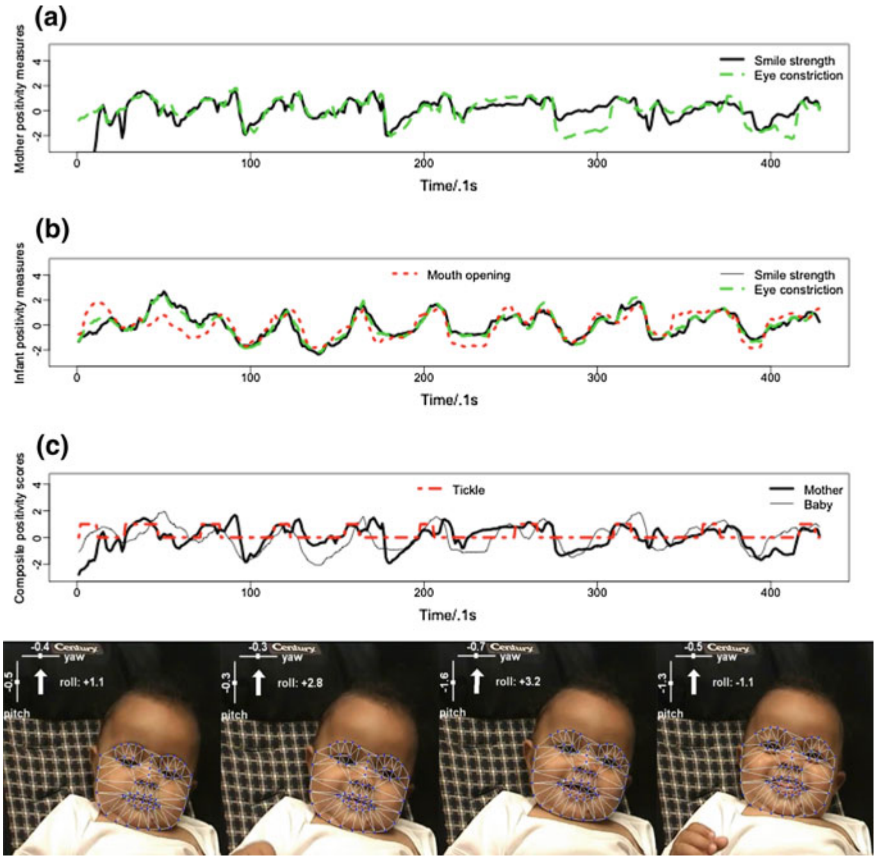


Fig. 17.1 **a**, **b** Plots of the automated measures used to indicate the dyad members' positivity levels and **c** time series of mother and infant composite positivity scores plotted with "tickle," a binary index indicating when the mother was tickling the baby. The *bottom* figure shows the AFA4's Active Appearance Model (AAM). The AAM is a mesh that track facial features over time (from *left* to *right*) while separately modeling rigid head motion (yaw, pitch, and roll, visible in the *upper left* hand portion of the images)

mother-infant interaction process (due, e.g., to changes or perturbations from the environment) that may otherwise bias the estimation of coupling dynamics between the dyad members.

We then compared Model 1 to Model 2, a model consisting of the linear effect of tickling, a smooth of the time trend in the data, and tensor products of the dyad members' lagged positivity levels and time to capture possible nonlinear intrinsic dynamics of each dyad member as well as the dyad member's association with the other dyad member as:

Model 2

$$\begin{aligned}\mu_{Mom,t} &= \beta_{M1} Tickle_t + f_{M1}(Time_t) + f_{M11}(Mom_{t-1}, Time_t)\\ &\quad + f_{M12}(Mom_{t-2}, Time_t) + f_{M13}(Infant_{t-1}, Time_t)\\ \mu_{Infant,t} &= \beta_{I1} Tickle_t + f_{I1}(Time_t) + f_{I11}(Infant_{t-1}, Time_t)\\ &\quad + f_{I12}(Infant_{t-2}, Time_t) + f_{I13}(Mom_{t-1}, Time_t)\end{aligned} \tag{17.3}$$

As distinct from standard linear autoregressive models such as that shown in (17.2), in this model, the autoregression effects for the mother [including the terms $f_{M11}(Mom_{t-1}, Time_t)$ and $f_{M12}(Mom_{t-2}, Time_t)$] and infant [including the terms $f_{I11}(Infant_{t-1}, Time_t)$ and $f_{I12}(Infant_{t-2}, Time_t)$] were allowed to be moderated by time in nonlinear ways. As such, these terms may be referred to as time-moderated autoregression effects. Of particular interest in this model were the time-moderated cross-regression effects [including the terms $f_{M13}(Infant_{t-1}, Time_t)$ and $f_{I13}(Mom_{t-1}, Time_t)$], which served as a proxy for delineating time-dependent fluctuations in mother and infant responsivity, respectively.

The two models proposed herein are estimated through the package MGCV in R via penalized least squares estimation. In practice, a variety of spline or penalized spline functions may be used to obtain the smoothed values [i.e., all terms involving *f*(.)] in these equations. Popular choices include cubic splines, *B*-splines, *P*-splines and other penalized regression splines (Green & Silverman, 1994). Here, we use the thin plate regression splines, which use an eigenvalue decomposition procedure to select piecewise regression spline coefficients that can maximize the amount of variance explained in the data. Thin plate regression splines have the advantages of (a) not having to choose knot locations, thereby reducing subjectivity in modeling and otherwise having optimal bases (Wood, 2006); and (b) being able to accommodate a higher number of predictors than other spline regression methods.

17.3 Results

The fit of Models 1 and 2 was compared using the Akaike information criterion (AIC; Akaike, 1973) and generalized cross-validation index (GCV), which can be taken as an estimate of the mean square prediction error based on a leave-one-out cross-validation estimation process, with lower values indicating better fit (Wood, 2006). Both the AIC and GCV values indicated that Model 2 with the tensor product terms provided better fit than the linear parametric Model 1 (see Table 17.1). To aid comparison, we include the parameter estimates, standard error estimates, and other output relevant for inferential purposes for both Models 1 and 2 in Table 17.1.

Key results from fitting Model 2 are depicted graphically in Figs. 17.2 and 17.3. The results indicated that six out of eight smooth (nonparametric) terms in Model 2, including five time-moderated tensor product terms, were found to be statistically

Table 17.1 Results from fitting Models 1–2 to the Dyad's data

	Model 1	Model 2
Parametric components	Parameter estimates (SE)	Parameter estimates (SE)
β_{M1}	−0.002 (0.03)	0.04 (0.02)
β_{M2}	1.29 (0.03)***	–
β_{M3}	−0.60 (0.03)***	–
β_{M4}	0.07 (0.04)	–
β_{I1}	0.06 (0.02)**	0.05 (0.02)**
β_{I2}	1.68 (0.04)***	–
β_{I3}	−0.71 (0.04)***	–
β_{I4}	−0.01 (0.01)	–
Nonparametric (smooth) components	edf	edf
$f_{M1}(time)$	46.93***	5.54
$f_{I1}(Time)$	1.00	8.87***
$f_{M11}(Mom_{t-1}, Time)$	–	57.02***
$f_{M12}(Mom_{t-2}, Time)$	–	35.77***
$f_{M13}(Infant_{t-1}, Time)$	–	21.07***
$f_{I11}(Infant_{t-1}, Time)$	–	2.94***
$f_{I12}(Infant_{t-2}, Time)$	–	5.76***
$f_{I13}(Mom_{t-1}, Time)$	–	3.86
Fit information		
Adjusted R^2	0.976	0.980
GCV	0.021	0.019
AIC	−887.79	−978.94

Note SE = Standard error estimates; edf = effective degrees of freedom; GCV = generalized cross-validation index; AIC = Akaike information criterion
***$p < 0.0001$; **$p < 0.001$; *$p < 0.01$

significant ($p < 0.0001$), with effective degrees of freedom (edfs) that deviated considerably from 1.0 (see second column of Table 17.1). An edf value that deviates substantially from 1.0 suggests that the associated smooth term is characterized by substantial deviations from linearity.[4] The smoothed time trend for the infant, plotted in Fig. 17.2a, indicated a relatively constant infant positivity for this dyad prior to $t = 200$, followed by a slight quadratic decline pattern that bounced back after $t = 300$. Such a relatively smoothed time trend was desirable from the perspective of modeling lagged dependencies between the mother and the infant because it helped remove trends that may have biased estimates of lagged

[4] Edfs are inversely related to the smoothing parameter used in the penalized basis functions to smooth out "wiggliness" in the data. Roughly speaking, they may be viewed as weights that map the penalized smoothed coefficient of a covariate to the unpenalized linear parametric coefficient associated with the covariate. An edf value that is close to zero implies that a particular covariate does not have statistically significant effect on the dependent variable whereas an edf value close to 1.0 suggests insufficient evidence for the effect of the covariate to be nonlinear.

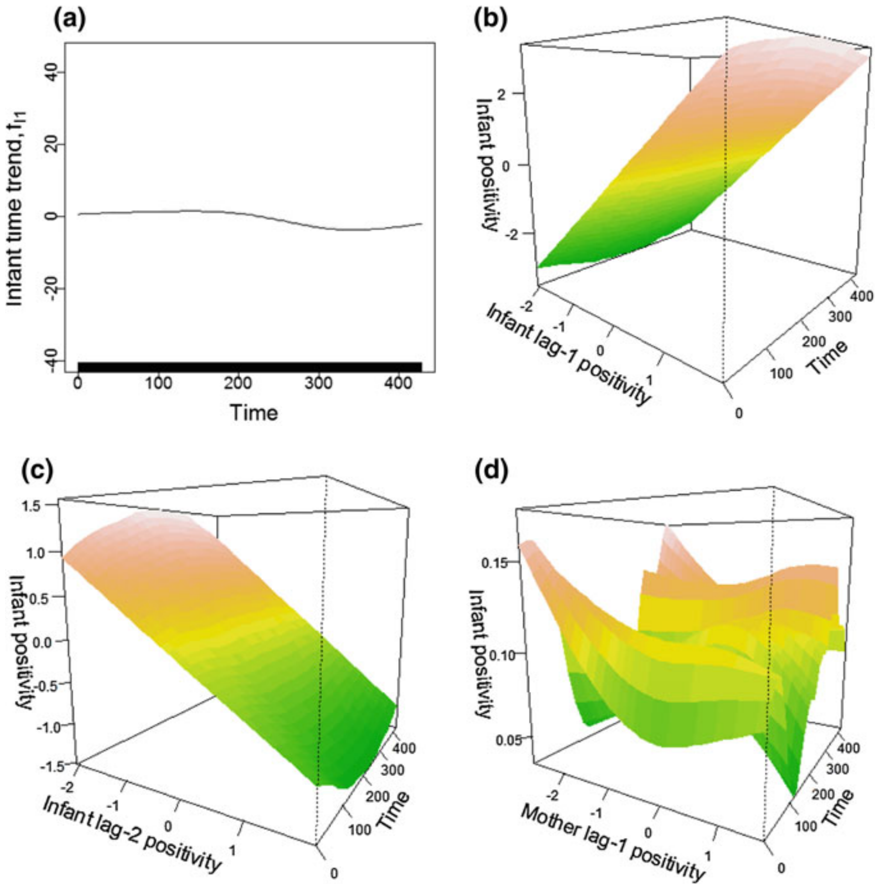


Fig. 17.2 Plots depicting selected effects on infant positivity from Model 2: **a** the smoothed time trend of the infant's positivity level; **b** the joint effect of infant's positivity level at time $t - 1$ and time on infant positivity at time t, or in other words, time-moderated lag-1 auto-regression effect of infant positivity; **c** the joint effect of infant's positivity level at time $t - 1$ and time on mother positivity and time t, or in other words, time-moderated mother responsivity; **d**–**f** rotated views of the joint effect of mother's positivity level at time $t - 1$ and time on infant positivity at time t, or in other words, time-moderated infant responsivity, from different angles; and **g**–**i** 2-dimensional slices of plots (**d**)–(**f**) at time t = 20, 165, and 380 with the values of other variables held at their respective medians

within- and between-person dependencies while not over-extracting more subtle micro-level changes. In this way, sufficient ebbs and flows were retained in the data to be accounted for by other independent variables. The smoothed time trend played a statistically significant role (i.e., the 95% confidence interval of this smoothed term generally did not include zero) in explaining the infant's but not the mother's positivity levels.

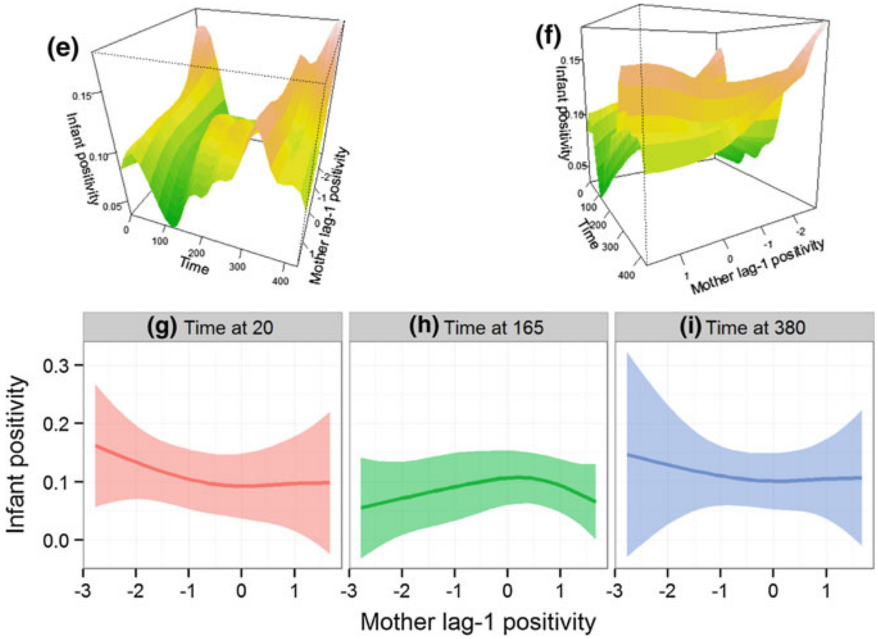


Fig. 17.2 (continued)

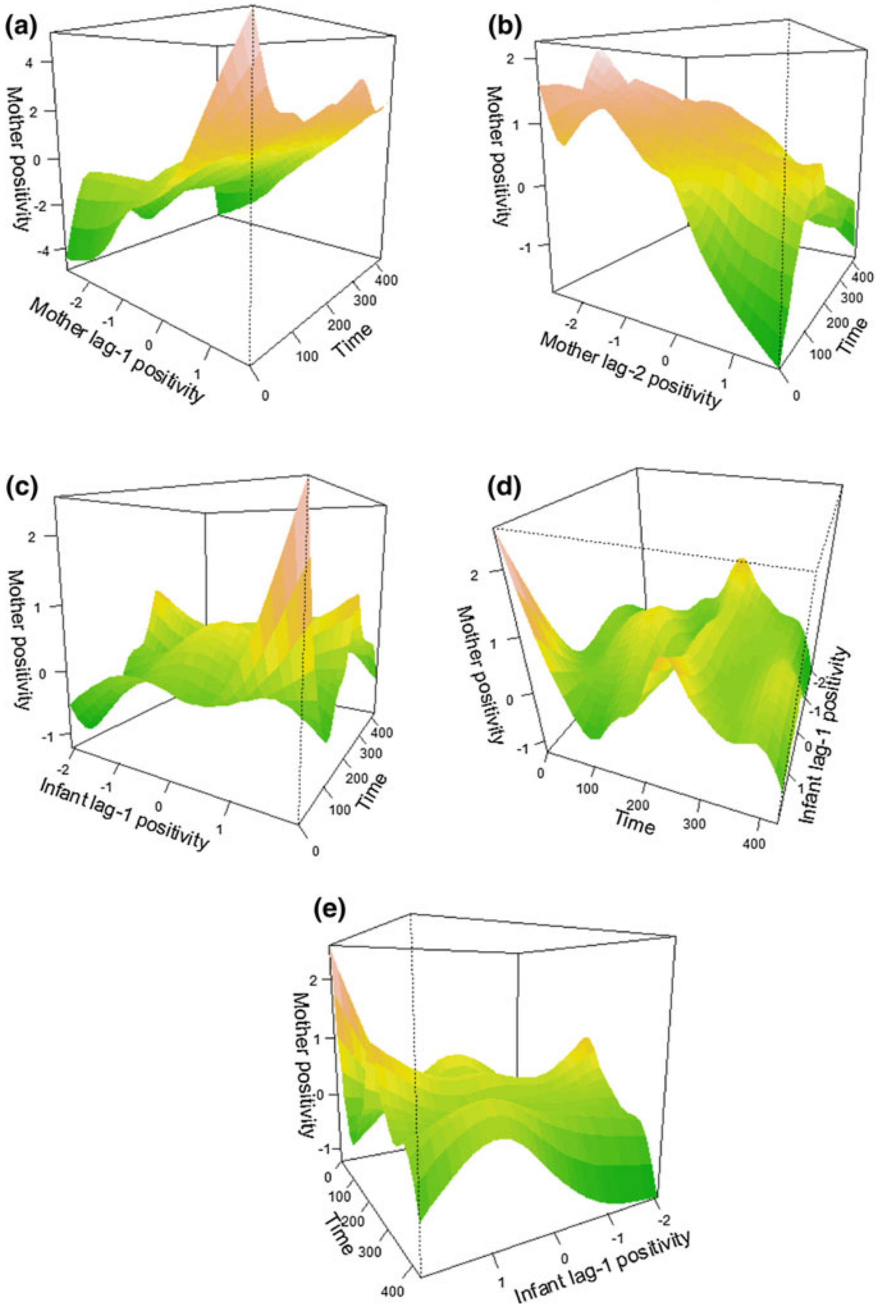
The smooth of the tensor product between time and the infant's lag-1 autoregressive effect (see Fig. 17.2b) revealed that the lag-1 autoregressive effect of infant positivity was positive and mostly linear. The corresponding smooth of the tensor product involving time and the infant's lag-2 autoregressive effect (see Fig. 17.2c) was generally negative. The angles of the planes depicting the associations between the infant's current and previous positivity at time $t-1$ (i.e., at lag 1 or 0.1 s ago) and $t-2$ (i.e., at lag 2 or 0.2 s ago) remained largely constant. Thus, these associations remained largely linear and only showed limited moderation by time, as confirmed by the relatively low edfs of these two smooths (2.94 and 5.76, respectively). In contrast, the mother's positivity at time $t-1$ was observed to show a positive exponential association with the infant's positivity at the beginning of the interaction episode that became linear and slightly attenuated toward the end of the interaction (see Fig. 17.3a). Similar to the infant, the lag-2 auto-regression effect of the mother's positivity was largely negative, but showed more over-time fluctuations compared to the infant (see Fig. 17.3b).

Forcing the time-moderated auto- and cross-regression effects in Model 2 to be time-invariant as in Model 1 obscured the statistically significant but fluctuating cross-regression relation from the infant to the mother as captured by f_{M13} $(Infant_{t-1}, Time)$. This smooth term was estimated to be statistically significant, whereas the reverse cross-regression relation from the mother to the infant, as captured by $f_{I13}(Mom_{t-1}, Time)$, was not statistically significant (see the second column of Table 17.1). This was in contrast to the results from fitting Model 1 with

Fig. 17.3 Plots depicting selected effects on the mother's positivity from Model 2: **a** the ▶ smoothed time trend of the infant's positivity level; **b** the joint effect of infant's positivity level at time $t - 1$ and time on infant positivity at time t, or in other words, time-moderated lag-1 auto-regression effect of infant positivity; **c** the joint effect of infant's positivity level at time $t - 1$ and time on mother positivity and time t, or in other words, time-moderated mother responsivity; **d–f** rotated views of the joint effect of mother's positivity level at time $t - 1$ and time on infant positivity at time t, or in other words, time-moderated infant responsivity, from different angles; and **g–i** 2-dimensional slices of plots (**d**)–(**f**) at time t = 20, 165, and 380 with the values of other variables held at their respective medians

time-invariant cross-regression effects, wherein the overall or time-invariant cross-regression effects, β_{I4} and β_{M4}, were both estimated to be not significantly different from zero (see the first column of Table 17.1) and would have led to the erroneous conclusion that there was no evidence for parental responsivity in this dyad.

To aid understanding of results from the GAMM, we rotated the plot of these cross-regression associations at different angles (see Figs. 17.2d–f and 17.3c–e, respectively), and depict these associations at particular slices of time (i.e., at t = 20, 165 and 380; see Figs. 17.2g–i and 17.3f–h) with their approximate 95% confidence intervals (obtained by adding and subtracting two standard errors from the predicted trajectories). Inspection of these plots revealed that the mother → infant coupling effect (infant responsivity) was fleeting and characterized by noisy fluctuations over time. This led to very wide confidence intervals that generally overlapped with zero. Instead, the mother's influence on the infant was manifested primarily through the mother's tickling action, which led to statistically significant concurrent elevations in the infant's but not the mother's positivity levels ($\beta_{I1} = 0.06, \mathrm{SE} = 0.02, p < 0.01$; compared to $\beta_{M1} = -0.002, SE = 0.03,\ n.s.$). In contrast, the statistically significant infant → mother coupling effect (parental responsivity) still varied over time, but in slightly more systematic ways and was characterized by tighter confidence bounds. This cross-regression effect was observed to be positive at the beginning of the interaction (at t = 20; see Fig. 17.3g), possibly reflecting the role of the infant in elevating the mother's subsequent positivity during this period. At later time points, this effect became negative, with slight quadratic trend at t = 165 and 380. Thus, at t = 165, for instance, low infant positivity (<0) at time $t - 1$ tended to drive the mother to increase her positivity (possibly in hopes of eliciting positivity from the infant) but this association was slightly negative or became attenuated at high (>0) infant positivity at time $t - 1$. The presence of a time trend, together with the over-time heterogeneities in the coupling dynamics described earlier, rendered the dyad's dynamics nonstationary. In addition, the continuous over-time changes in the infant → mother cross-regression effect also provided a glimpse into how the dyad self-organized into distinct states over the course of a brief interaction, including periods where the mother's positivity was reinforced and strengthened by the infant's positivity, as well as periods where the mother's positivity was inversely related to the infant's past positivity levels. The rich dynamics embedded in such within-dyad fluctuations could have been easily bypassed if models assuming stationarity were used.


(a)
Mother positivity
4
2
0
-2
-4
Mother lag-1 positivity
-2
-1
0
1
Time
0
100
200
300
400
(b)
Mother positivity
2
1
0
-1
Mother lag-2 positivity
-2
-1
0
1
Time
0
100
200
300
400
(c)
Mother positivity
2
1
0
-1
Infant lag-1 positivity
-2
-1
0
1
Time
0
100
200
300
400
(d)
Mother positivity
2
1
0
-1
Time
0
100
200
300
400
Infant lag-1 positivity
-2
-1
0
1
(e)
Mother positivity
2
1
0
-1
Time
0
100
200
300
400
Infant lag-1 positivity
1
0
-1
-2

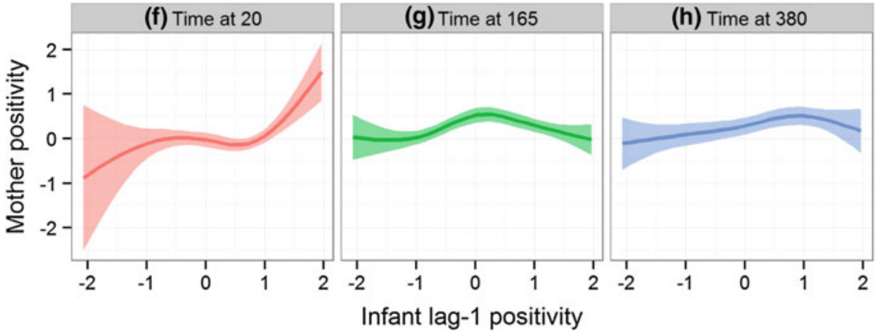


Fig. 17.3 (continued)

17.4 Discussion

Research on infants' interactive dynamics in the past few decades has emphasized static, group-based notions of development. However, there may be a growing consensus that change, as opposed to stability, is the operating norm of children's socioemotional functioning and development (Fogel & Thelen, 1987), both between individuals and between behaviors (de Weerth & van Geert, 2002). Using the current dynamic modeling approach, we found that the dynamics of a mother-infant dyad changed substantially within a relatively brief interaction episode, and also differently between members of the infant-mother dyad.

The current work used the data from one particular dyad as an illustrative example. While such an idiographic approach is a critical step toward understanding human dynamics (Molenaar, 2004), the MGCV framework can readily accommodate data from multiple subjects, thus providing some flexibility in capturing between-dyad differences in the form of random effects. Ultimately, the question of whether homogeneous measurement and change structures may be assumed to justify pooling multiple subjects' (or dyads') time series deserves careful consideration (Hamaker, Dolan, & Molenaar, 2005).

In our empirical analysis, we used composite scores from mean aggregation to eliminate some of the interindividual differences in localized facial dynamics over time. Other approaches, such as one that utilizes explicit parametric or nonparametric measurement models to link each of the automated facial measures to the underlying constructs of infant and mother positivity, may also be possible. In addition, the generalized additive framework implemented in MGCV can accommodate measurement responses that are members of the exponential family. In the present article, we have focused on modeling continuous data. Other kinds of measurement functions may also be assumed and should be utilized where appropriate (Moustaki, 2000).

We used one particular option within the MGCV library, thin-plate spline regression data (Wang, Du, & Shen, 2013; Wood, 2003, 2006), which can be used to build nonlinearity nonparametrically into the model while automatically selecting the placement of knot points in obtaining piecewise smooths of the data. Despite the practical advantages of this approach, caution should still be exercised, given that the final number of selected knot points and the corresponding edfs may still be sensitive to the starting values specified by the user.

Our illustrative application is but one example of how sources of nonstationarities and heterogeneities in dynamics can be decomposed to shed light on the change phenomenon of interest. Our hope is that the illustration helps demonstrate how spline and nonparametric functions may be utilized to aid the development and exploration of dynamic systems models, particularly models that are characterized by self-organizing properties, or other sources of nonstationarities and heterogeneities in dynamics. On a practical note, these models open new ground in the analysis of group dynamics. By explicitly modeling time-varying changes in both autoregression and interactive (cross-regression) parameters, our use of GAMM revealed time-varying dyadic processes obscured by less flexible time-invariant models. At the same time, this class of models can reveal slow changes in individual process that may masquerade as interperson influence.

Acknowledgements The authors thank the families participating in the research and acknowledge support from NIH grant R01GM105004, NSF grant BCS-0826844, Penn State Quantitative Social Sciences Initiative and UL TR000127 from the National Center for Advancing Translational Sciences.

References

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Oxford, England: Lawrence Erlbaum.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium in Information Theory* (pp. 267–281). Tsahkadsor, Armenia: Akad.Kiadó).

Barton, Scott. (1994). Chaos, self-organization, and psychology. *American Psychologist, 49*(1), 5–14.

Brazelton, T. B., Koslowski, B., & Main, M. (1974). The origins of reciprocity: The early mother-infant interaction. In M. Lewis & L. Rosenblum (Eds.), *The effects of the infant on its caregiver* (pp. 137–154). New York, NY: Wiley-Interscience).

Chatfield, C. (2004). *The analysis of time series: An introduction* (6th ed.). Boca Raton, FL: CRC Press.

Chow, S.-M., Haltigan, J. D., & Messinger, D. S. (2010). Dynamic patterns of infant-parent interactions during Face-to-Face and Still-Face episodes. *Emotion, 10*(1), 101–114.

Chow, S.-M., Hamaker, E. J., & Allaire, Jason C. (2009). Using innovative outliers to detecting discrete shifts in dynamics in group-based state-space models. *Multivariate Behavioral Research, 44,* 465–496.

Chow, S.-M., Zu, J., Shifren, K., & Zhang, G. (2011). Dynamic factor analysis models with time-varying parameters. *Multivariate Behavioral Research, 46*(2), 303–339.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis in the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

De Jong, P., & Penzer, J. (1998). Diagnosing shocks in time series. *Journal of the American Statistical Association, 93,* 796–806.

de Weerth, C., & van Geert, P. (2002). Changing patterns of infant behavior and mother-infant interaction: Intra and interindividual variability. *Infant Behavior and Development, 24,* 347–371.

Del Negro, M., & Otrok, C. (2008). Dynamic factor models with time-varying parameters: Measuring changes in international business cycles. *Federal Reserve Bank of New York Staff Reports, 326,* 1–46.

Durbin, J., & Koopman, S. J. (2001). *Time series analysis by state space methods*. New York, NY: Oxford University Press.

Ekas, N., Haltigan, J. D., & Messinger, D. S. (2013). The dynamic still-face effect: Do infants decrease bidding over time when parents are not responsive? *Developmental Psychology, 49*(6), 1027–1035.

Ekman, P., & Friesen, W. V. (1978). *Manual for the facial action coding system*. Palo Alto, CA: Consulting Psychologists Press.

Feldman, R., & Greenbaum, C. W. (1997). Affect regulation and synchrony in mother-infant play as precursors to the development of symbolic competence. *Infant Mental Health Journal, 18*(1), 4–23.

Feldman, R., Greenbaum, C. W., Yirmiya, N., & Mayes, L. C. (1996). Relations between cyclicity and regulation in mother-infant interaction at 3 and 9 months and cognition at 2 years. *Journal of Applied Developmental Psychology, 17*(3), 347–365.

Fogel, A., & Thelen, E. (1987). Development of early expressive and communicative action: Reinterpreting the evidence from a dynamic systems perspective. *Developmental Psychology, 23,* 747–761.

Green, P. J., & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*. Boca Raton, FL: CRC Press.

Haken, H. (1977/83). Synergetics, and introduction: Non-equilibrium phase transitions and self-organization in physics, chemistry and biology. Berlin, Germany: Springer.

Halpin, P. F., & De Boeck, P. (2013). Modelling dyadic interaction with Hawkes processes. *Psychometrika, 78,* 793–814. doi:10.1007/s11336-013-9329-1.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2005). Statistical modeling of the individual: Rationale and application of multivariate stationary time series analysis. *Multivariate Behavioral Research, 40,* 207–233.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological), 55*(4), 757–796.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London, UK: Chapman and Hall.

Henderson, R. K., Snyder, H. R., Gupta, T., & Banich, M. T. (2012). When does stress help or harm? The effects of stress controllability and subjective stress response on stroop performance. *Frontiers in Psychology, 3,* 484–498.

Heywood, I., Cornelius, S., & Carver, S. (2006). *An introduction to geographical information systems* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Ibrahim, A., Leelahanon, S., & Li, Q. I. (2005). Efficient estimation of a semiparametric partially linear varying coefficient model. *Annals of Statistics, 33*(1), 258–283.

Isabella, R. A., & Belsky, J. (1991). Interactional synchrony and the origins of infant-mother attachment: A replication study. *Child Development, 62*(2), 373–384.

Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., Jasnow, M. D, Rochat, P., et al. (2001). Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the Society for Research in Child Development, 66*(2), i-viii, 1–132.

James, G. (2002). Generalized linear models with functional predictor variables. *Journal of the Royal Statistical Society: Series B, 64,* 411–432.

Kelso, Scott J. A. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Kochanska, G., Forman, D. R., & Coy, K. C. (1999). Implications of the mother-child relationship in infancy for socialization in the second year of life. *Infant Behavior and Development, 22*(2), 249–265.

Landry, S. H., Smith, K. E., Miller-Loncar, C. L., & Swank, P. R. (1997). Predicting cognitive-language and social growth curves from early maternal behaviors in children at varying degrees of biological risk. *Developmental Psychology, 33*(6), 1040–1053.

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research, 32,* 215–253.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, UK: Chapman and Hall.

McKeown, G. J., & Sneddon, I. (2014). Modeling continuous self-report measures of perceived emotion using generalized additive mixed models. *Psychological Methods, 19*(1), 155–174.

Messinger, D. S., Mahoor, M. H., Chow, S.-M., & Cohn, J. F. (2009). Automated measurement of facial expression in infant-mother interaction: A pilot study. *Infancy, 14*(3), 285–305.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., et al. (2014). *Psychometric considerations in game-based assessment*. Redwood City, CA: Authors. Printed by CreateSpace Independent Publishing Platform.

Molenaar, P. C. M. (1994). Dynamic latent variable models in developmental psychology. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 155–180). Thousand Oaks, CA: Sage Publications.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific pyschology-this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2*, 201–218.

Moustaki, I. (2000). A latent variable model for ordinal variables. *Applied Psychological Measurement, 24,* 211–223.

Pagan, A. (1980). Some identification and estimation results for regression models with stochastically varying coefficients. *Journal of Econometrics, 13,* 341–363.

Smith, Linda B., & Thelen, Esther. (1993). *A dynamic systems approach to development*. Cambridge, MA: MIT Press.

Soller, A., & Stevens, R. H. (2007). Applications of stochastic analyses for collaborative learning and cognitive assessment. In G. Hancock & K. Samuelson (Eds.), *Advances in latent variable mixture models*. Greenwich, CT: Information Age Publishing.

Stern, D. N. (1985). *The interpersonal world of the infant: A view from psychoanalysis and developmental psychology*. New York, NY: Basic Books.

Stock, J. H., & Watson, M. H. (2008). Forecasting in dynamic factor models subject to structural instability. In J. Castle & N. Shephard (Eds.), *The methodology and practice of econometrics, a festschrift in honour of Professor David F. Hendry*. Oxford, England: Oxford University Press.

Tarvainen, M. P., Georgiadis, S. D., Ranta-aho, P. O., & Karjalainen, P. A. (2006). Time-varying analysis of heart rate variability signals with Kalman smoother algorithm. *Physiological Measurement, 27,* 225–239.

Thelen, E. (1989). Self-organization in developmental processes: Can systems approaches work? In M. R. Gunnar & E. Thelen (Eds.), *Systems and development* (pp. 77–117). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tronick, E. (1989). Emotions and emotional communication in infants. *American Psychologist, 44,* 112–119.

Tronick, E. Z., & Gianino, A. (1986). Interactive mismatch and repair: Challenges to the coping infant. *Zero to Three Bulletin of the National Center for Clinical Infant Programs, 3,* 1–6.

Wang, X., Du, P., & Shen, J. (2013). Smoothing splines with varying smoothing parameter. *Biometrika, 100*(4), 955–970.

Weiss, A. A. (1985). The stability of the AR(1) process with an AR(1) coefficient. *Journal of Time Series Analysis, 6,* 181–186.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(1), 95–114.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Raton, FL: CRC Press.

Chapter 18
Modeling Collaboration with Social Network Models

Tracy M. Sweet

Abstract Social network models are used to infer relationships about networks, network structures, and other attributes. Many network models focus on inference about a single network and relate node- or tie-level covariates to ties. When network data involve multiple, independent networks, another type of network model is used that both generalizes the findings of single-network models and infers relationships at higher levels in the model, addressing new research questions. In this chapter, we present commonly used network models in education research and describe how these models can be used to inform research on collaboration and team dynamics.

Keywords Social network analysis • Social network models • Latent space models • Mixed membership models • Advice-seeking, multilevel, collaboration

18.1 Introduction

A *social network* is defined as the set of relationships among a group of individuals, such as friendship among a group of students. Social networks are studied in a number of disciplines and appear naturally in the social sciences because fields such as education, political science, and sociology involve individuals interacting with one another. Examples of social networks include friendship, collaborations among researchers, advice seeking among teachers, and trade alliances among countries, but any type of interaction could be characterized as a network. For example, during a collaborative task or group project, individuals are interacting with one another, often with different frequencies and for different purposes, and we could construct various networks from these interactions.

T.M. Sweet (✉)
University of Maryland, College Park, MD, USA
e-mail: tsweet@umd.edu

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_18

Social network analysis refers to the set of quantitative methods used for relational data, data collected about the network ties and the individuals in the network. *Social network models* are a subset of these methods that focus on statistical inference. The network—the set of ties among all individuals—is considered to be the outcome or dependent variable, and we then employ a network model to estimate the associations of various network features with the network ties. Network models can be subdivided into two broad classes of models: (a) those that model the entire network and (b) those that model the probability of a tie. We focus on the latter in this chapter.

Social network models for a single social network generally relate network structures or nodal covariates to the set of network ties. Thus much of the analysis occurs at the node level; that is, we are interested in which node attributes are associated with network ties. For example, we might be interested in whether students of the same race are more likely to be friends than students of different races. In education, however, we often study more than one network at a time. We collect data on friendship in multiple classrooms of students or collaboration among teachers in multiple schools, and we are interested in the broad patterns that occur across all schools. Does race play a factor in all schools, or only in schools with a small proportion of minority students? The same may be true for teams research; we are interested in the effects common to most teams. For example, one might be interested in how the network among team members is related to team performance or how the goal of the team or makeup of the team is related to certain relationships that develop among team members. For these and other questions relating the network to network-level outcomes, hierarchical social network models are required.

Hierarchical social network models accommodate multiple, isolated networks, and these models assume that networks are independent of one another. For example, we might study networks of classrooms of students and collect network data on multiple classrooms, or we might study team interactions and collect data on multiple teams. In many situations, we can assume that these teams do not have ties to one another. The term *hierarchical* is borrowed from the hierarchical linear modeling literature (Raudenbush & Bryk, 2002) for nested data. Multiple social networks are also an example of clustered data; a set of network ties is nested within each network. Similarly, the term *multilevel network models* is also sometimes used to describe these models.

The purpose of this chapter is to introduce social network models, highlighting hierarchical network models used in education research that also can inform collaboration research. The remainder of the chapter is organized in the following way. We begin with an introduction to social network analysis and a summary of common statistical models. We then present two hierarchical social network models and provide examples to illustrate the types of research questions that can be addressed with these models.

18.1.1 Terminology

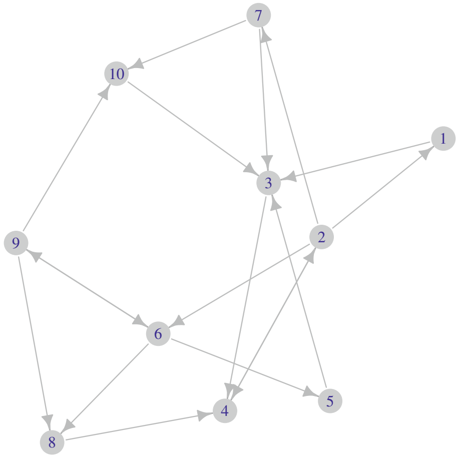
There are several ways to represent a social network. One common way is with an *adjacency matrix*: a square matrix Y with dimension equal to the number of nodes such that the entry Y_{ij} is the value of the relationship from individual i to individual j. Often these relationship values are binary—the presence or absence of said relationship—but they need not be. Friendship ties, for example, may be ordinal or continuous and represent a measure of closeness or frequency of interaction.

We also use figures to visually represent social networks, and the most common figure is a *sociogram*, in which the nodes or individuals are depicted as vertices and the ties are plotted as edges. For binary ties, one would observe the presence or absence of edges, whereas edge color or width may represent valued ties, and arrows convey directed (or asymmetric) relationships. Figure 18.1 shows an example of a binary, directed network with 10 individuals.

Methods for network analysis generally fall into two categories: (a) *exploratory analysis* and (b) *inferential analysis*. Although these classes are not mutually exclusive, the latter generally focuses on statistical models. Exploratory or descriptive statistics include summary measures at both the network and node levels. The most common network-level statistic is *density*, defined as the proportion of observed (binary) ties out of all possible ties. Note that there are $n(n-1)$ pairs of individuals in the network and thus $n(n-1)$ possible ties. Networks that are sparse have densities near 0, and networks that are dense have densities closer to 1, although what constitutes a dense network is context specific.

Other network statistics include measures about the nodes; *in-degree* is the number of ties that a node receives, whereas *out-degree* is the number of ties that a node sends. If ties are undirected, we use the term *degree*, which is the number of ties that a given node has. These measures can identify particularly influential nodes or nodes that are more isolated. The distribution of (in/out)-degree is often

Fig. 18.1 Visual representation of a social network where vertices represent individuals and the presence of each directional relationship is indicated by an arrow



important to characterize the network. Other network statistics focus on network ties to determine central ties/edges, and still others focus on certain network features or patterns, such as the number of reciprocated ties or triads. For a comprehensive list and other information, see Wasserman and Faust (1994) or Kolaczyk (2009).

For statistical inference, we often employ a model. Social network statistical models generally treat the entire network—the full set of absent and present network ties—as the dependent variable and then estimate the effects of node attributes or network features. There are a variety of network models. *Exponential random graph models* (Wasserman & Pattison, 1996) model the probability of observing a particular network out of the space of all possible networks. In fact, Zhu (Chap. 19) uses these models to explore team assembly in Chap. X. *Latent space models* (Hoff, Raftery, & Handcock, 2002) and *stochastic blockmodels* (Holland, Laskey, & Leinhardt, 1983) model the probability of a tie (or tie value) as based on some latent structure. Because the ties are modeled as independent given this latent structure, we use the term *conditionally independent tie models* to describe these models.

There are also a number of extensions to single-network models. A set of individuals whose relationships are collected across several time points requires a longitudinal network model. The most common class of longitudinal model is the *stochastic actor-oriented model* (Snijders, 1996), which models the dynamic process of ties changing, forming, or persisting over time. Schector and Contractor (Chap. 14) employ *relational event models*, which are a different type of dynamic model where the tie formation sequence is of interest, and this work is not unrelated to the *point process models* that Halpin and von Davier introduce in Chap. 15.

As noted in Sect. 18.1, another extension is to multiple networks, and hierarchical network models (Sweet, Thomas, & Junker, 2013) describe a framework for extending single-network models to accommodate multiple networks. Unlike in longitudinal models, the networks do not include the same nodes and are separate, independent networks, such as teachers within schools or individuals on teams. We are interested in analyzing these networks together because we have reason to believe that these networks are similar and that analyzing across multiple sites improves generalizability. Note, however, that the term *multilevel model* in the social network literature generally refers to models for multilevel networks that involve ties between different levels (Wang, Robins, Pattison, & Lazega, 2013). For example, for given individuals clustered in organizations, we could imagine ties among the individuals, ties from each individual to one or more organizations, and ties across organizations.

18.2 Conditionally Independent Social Network Models

In many statistical models, we assume that observations are independent. For example, in a simple linear regression relating height and weight, we assume that the data are collected from independent individuals. A linear regression model would not be appropriate for data collected from one family because their weights are unlikely to be independent of each other. Other models are then necessary to accommodate the lack of independence. For example, there are longitudinal models for repeated measures of the same individual or multilevel models to accommodate higher rates of correlation among individuals in the same group or cluster.

If we consider network ties as our observations, we can immediately see that ties in a network are rarely independent. For example, individual *i*'s relationships to individuals *j* and *k* affect the relationship between *j* and *k* because *j* and *k* are likely to interact through *i*. In addition, whether *i* has ties with *j* and *k* also influences ties with ℓ and *m*, because individuals have a finite amount of time for social interactions. Similarly, the absence of ties is also related. If *i* and *j* are friends, and *j* does not have a tie to *k*, it is then less likely that *i* will have a tie with *k*; one could imagine progressively more complex examples.

Thus the dependence structure among network ties is very complex and quite difficult to formalize. One way is to explicitly define dependence in the model by including terms in the model. For example, one could incorporate parameters that measure certain structures, for example, the proportion of ties that are reciprocated (a tie from *i* to *j* also appears with a tie from *i* to *j*) or the number of adjacent ties that form a triad (a tie from *i* to *j* and *j* to *k* also appears with a tie from *i* to *k* or *k* to *i*). Any structure can be included, but the assumption is that ties are independent conditional on these structures, which makes choosing which network measures to include an important and generally difficult issue.

Another way to accommodate tie interdependence is to include latent variables that represent this structure. The models that take this approach are called *conditionally independent network models*, and they make the assumption that ties or dyads are independent conditional on other parameters or latent variables in the model. These models are also useful because once these latent variables are estimated, ties can be modeled as independent, which facilitates model estimation.

We review two such network models: latent space models (LSM; Hoff et al., 2002) and stochastic block models (Holland et al., 1983). Latent space models assume that the individuals in a network occupy a position in a latent social space; for binary ties, individuals far apart in this social space are unlikely to have a tie, and individuals very close in this space are much more likely to have a tie. The probability of a tie is then a function of the distance between the pair of nodes in this space. This heuristic can be modified for ordinal or continuous ties as well. Given these positions and any other covariates in the model, the ties are assumed to be independent.

Given binary network *Y*, we define $Y_{ij} = 1$ as a tie from *i* to *j*. A simple LSM is given as

$$\begin{aligned} P(Y|X, Z, \beta) &= \prod_{ij} P(Y_{ij}|X_{ij}, Z_i, Z_j, \beta) \\ \operatorname{logit}[P(Y_{ij} = 1)|X_{ij}, Z_i, Z_j, \beta] &= \beta X_{ij} - |Z_i - Z_j|, \end{aligned} \tag{18.1}$$

where $Z = (Z_1, \ldots, Z_n)$ such that Z_i is the latent space position for node i and Z_j is the position for j. We also include covariates X, which can include both node-level covariates, such as age or gender, and dyad-level covariates, such as difference in age or being of the same gender.

LSMs have been used in education research to estimate the effects of various covariates on network ties. One can also estimate the latent positions, which could inform substantive research. For example, we might be interested in whether the network has any subgroup structure or other interesting structure not visible in a network plot.

If we have reasons to believe that individuals in the network self-group, stochastic blockmodels may be preferable. These models assume that individuals belong to a latent cluster or block and that block membership then determines the likelihood (or value) of a tie. We generally assume that ties within blocks are more likely than ties between blocks—as is the case for self-grouping—but the model can accommodate other assumptions as well.

For binary ties, a stochastic blockmodel is given as

$$P(Y|g, B) = \prod_{ij} P(Y_{ij}|g_i, g_j, B) P(Y_{ij} = 1|g_i, g_j, B) = B_{g_i g_j}, \tag{18.2}$$

where g_i is the block membership for node i, g_j is membership for j, and B is a matrix of block–block tie probabilities. The B matrix has dimension equal to the number of blocks, which is specified a priori. This model is deceptively simple, but in fact, g_i is estimated for all nodes in the network, and all entries of the B matrix are also estimated. Blockmodels are useful for identifying the cluster membership for each node, and this could be particularly informative for studies where individuals have assigned roles or are organized into subgroups already.

A common extension to this model is the mixed membership stochastic blockmodel (MMSBM; Airoldi, Blei, Fienberg, & Xing, 2008), in which the block membership probability varies for each node. In fact, g_i is replaced by θ_i, a probability vector for belonging to each block. Thus block membership varies for each i and j when they interact in the network. The MMSBM is given as

$$P(Y|B, S, R) = \prod_{ij} P(Y_{ij}|S_{ij}, R_{ji}, B) P(Y_{ij} = 1|S_{ij}, R_{ji}, B) = S'_{ij} B R_{ji}, \tag{18.3}$$

where S_{ij} is the block membership indicator for i when sending a tie to j and R_{ji} is the block membership indicator for j when receiving a tie from i. Note that S_{ij} and R_{ji} vary for all combinations of (i,j) and are determined by the block probability vector θ_i. In fact, one way to illustrate how a network is generated from this model is to write the MMSBM as a hierarchical Bayesian model:

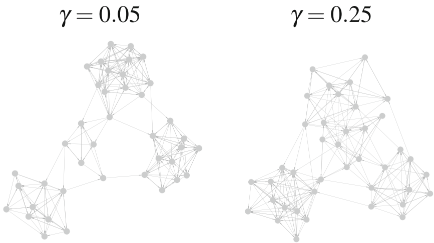


Fig. 18.2 Examples of three networks generated from a mixed membership stochastic blockmodel with different γ parameters. When γ is near 0, subgroups become very insular, with few ties across blocks

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}\left(S_{ij}^{'} B R_{ji}\right), \\ S_{ij} &\sim \text{Multinomial}\left(\theta_i, 1\right), \\ R_{ji} &\sim \text{Multinomial}\left(\theta_j, 1\right), \\ \theta_i &\sim \text{Dirichlet}\left(\gamma/g\right), \\ B_{\ell m} &\sim \text{Beta}\left(a_{\ell m}, b_{\ell m}\right), \\ \gamma &\sim \text{Gamma}\left(c, d\right), \end{aligned} \tag{18.4}$$

where g is the number of subgroups or blocks assumed to exist in the network. Written in this way, one can see how S_{ij} and R_{ji} vary for each pair of nodes. In addition, we include prior distributions for θ and γ, which represent additional layers of hierarchy in the model.

One advantage of the MMSBM is that it allows for nodes to belong to multiple groups so that it can be used when subgroup structure is less obvious. In fact, there is even a parameter γ that measures subgroup insularity (Sweet & Zheng, 2016, 2017). When γ is very small, θ is likely to be extreme such that each node has very high probability of belonging to one block and very little probability of belonging to the other blocks. As γ increases, θ becomes less extreme.

Consider the networks shown in Fig. 18.2. These networks are generated from a MMSBM with different values of γ. When γ is close to 0, we find networks that are quite insular, but when γ is larger, networks are much more integrated. Thus γ can estimate the amount of block or subgroup insularity (Sweet & Zheng, 2016, 2017).

18.3 Hierarchical Network Models

Many research questions cannot be adequately answered by a single network, namely, questions about the generalizability of these results to other networks as well as comparisons among networks. For example, a single network can tell you whether the network nodes tend to group by gender or if the more experienced teachers tend to be sought for advice, but we would need to examine several (if not many) networks to determine whether these relationships are specific to this network or if we find similar patterns across many networks and would expect these patterns generally. Furthermore, we may also be interested in network-level effects. For example, are dense networks associated with higher student achievement or less variability in instructional methods? To address these types of questions, we need to collect data across several networks, and in this framework, we consider each network as a separate (and independent) observation.

Thus we need a social network model that accommodates many networks. In a single-network model, the set of network ties is the outcome; now, with multiple networks, we have multiple sets of network ties clustered within each network. Recall that this nested structure is similar to the hierarchical linear model (Raudenbush & Bryk, 2002) structure of students clustered within classrooms and/or schools, and Sweet et al. (2013) introduced a framework called hierarchical network models for social network statistical models in which ties are clustered within networks. Most multilevel social networks for independent networks fall under this class of models. Similar to hierarchical linear models, these models allow covariate effects to vary across networks as well as estimates of network-level effects.

The general hierarchical network model is defined as

$$P(\mathrm{Y}|\mathrm{X},\Theta) = \prod_k P(Y_k|X_k\Theta_k)(\Theta_1,\ldots\Theta_k) \sim F(W_1,\ldots,W_k|\psi), \tag{18.5}$$

where Y_k is network k and X_k and Θ_k are the covariates and parameters for network k. Thus we can model the networks as independent replications conditional on the parameters in the model, which we model as having some joint distribution F. Note here that F can be any distribution, allowing for a variety of dependence assumptions.

We now consider two example hierarchical network models, the hierarchical latent space model and the hierarchical mixed membership stochastic block model, which are the respective extensions to the latent space model and MMSBM presented in Sect. 18.2.

18.3.1 Hierarchical Latent Space Models

The hierarchical latent space model (HLSM; Sweet et al., 2013) is a multilevel extension of the latent space model (Hoff et al., 2002) shown in Eq. (18.1). For a set of binary networks $Y_1, \ldots, Y_k$), a HLSM is given as

$$\begin{aligned} P(\mathrm{Y}|\mathrm{X},\mathrm{Z},\beta) &= \prod_k P(Y_k|X_k, Z_k, \beta_k), \\ \operatorname{logit}[P(Y_{ijk}=1)] &= \beta_k X_{ijk} - |Z_{ik} - Z_{jk}|, \\ \beta_k &\sim N(\mu_\beta, \sigma_\beta^2), \\ Z &\sim MVN(\mu_Z, \Sigma_Z), \end{aligned} \tag{18.6}$$

where Z_k is the set of latent space positions for network k and β_k includes a collection of parameters that are fixed or vary across networks. We usually write these models in a Bayesian framework so that we may include prior distributions for μ_β, μ_Z, σ_β^2, and Σ_Z.

HLSMs can be used to estimate the effects of a covariate across multiple networks as well as to estimate network-level effects. For example, Hopkins, Lowenhaupt, and Sweet (2015) fit HLSMs to estimate the aggregate effect of being an English language learner teacher on advice-seeking ties with other teachers and found that these teachers were less likely to seek advice and be sought for advice. Sweet and Zheng (2015) examined network-level effects of HLSMs using teacher advice-seeking ties and found that teachers seek advice regarding literacy with higher probability than they seek advice about mathematics.

HLSMs can also be used to examine how covariate effects vary across networks. To illustrate this, we fit HLSMs in which node-level covariate effects vary across networks. The data we used come from Pitts and Spillane (2009) and include teacher advice-seeking ties and demographic information from teachers in 15 elementary schools in one district. These schools are mostly elementary schools, but some schools are K − 8. Note that these are the same data analyzed by Sweet and Zheng (2015). The advice-seeking networks are shown in Fig. 18.3. Schools vary in both size and network density.

To examine how covariate effects vary across schools, we selected three binary tie-level variables: (a) whether the teachers have been teaching for a similar number of years, (b) whether the teachers have similar beliefs about innovative instruction, and (c) whether the teachers teach the same grade. We fit the following HLSM to the 15 teacher advice-seeking networks:

$$\begin{aligned} \operatorname{logit}[P(Y_{ijk}=1)] = \beta_{0k} + \beta_{1k}X_{1ijk} + \cdots + \beta_{3k}X_{3ijk} - |Z_{ik} - Z_{jk}|, \\ \beta_{ik} \sim N(\mu_i, \sigma_i^2), \\ \mu_i \sim N(0,1), \\ \sigma_i \sim \mathrm{Inv} - \mathrm{Gamma}(10, 15), \\ Z_{ik} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}\right). \end{aligned} \tag{18.7}$$

To fit this model, we used a Markov chain Monte Carlo (MCMC; Gelman, Carlin, Stern, & Rubin, 2014) algorithm coded in R (R Development Core Team, 2016), which can be accessed through the R package HLSM (Adhikari, Junker,

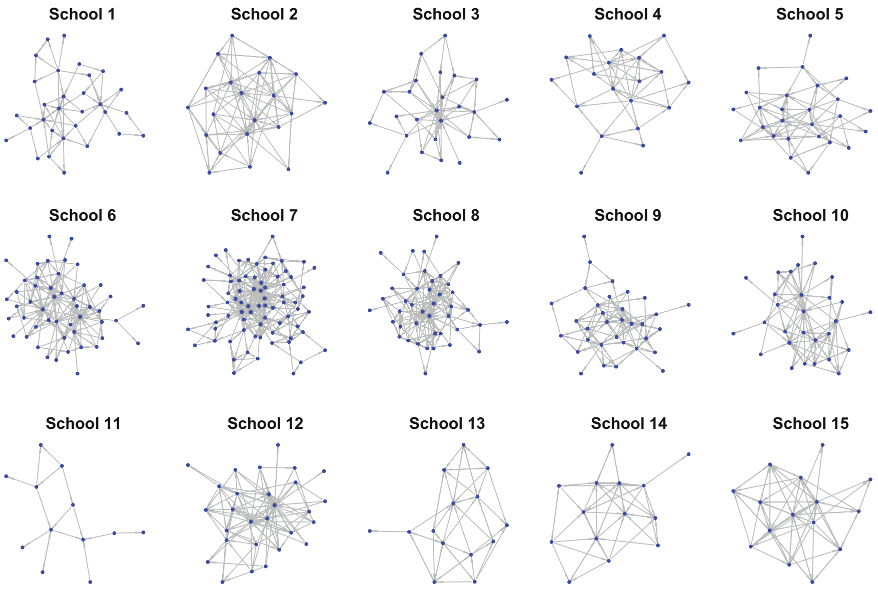


Fig. 18.3 Advice-seeking networks among teachers in 15 schools in one school district suggesting that advice seeking varies by school

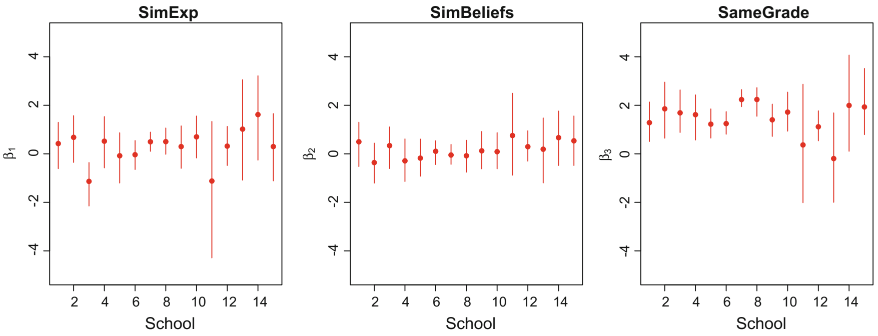


Fig. 18.4 The posterior mean (*point*) and 95% equal-tailed credible interval (*line*) are plotted for each school. Each plot shows the posterior summaries of a covariate effect across all 15 schools

Sweet, & Thomas, 2014). Models fit using MCMC result in samples from a Bayesian posterior distribution for each parameter. These distributions are generally summarized by their posterior means (or modes) and 95% credible intervals. The posterior means and corresponding credible intervals are plotted in Fig. 18.4.

The effect of having similar experience seems to be slightly (but not significantly) positive and varies across the schools. In fact, Schools 3 and 11 both suggest a negative impact of similar teaching experience. The credible intervals vary in width as well: Schools 11, 13, and 14 all have much wider intervals, suggesting less information in those schools. With respect to similar beliefs, we see very little effect on advice-seeking ties and little variability. Finally, there is a more positive and perhaps even significant effect of teaching the same grade, because the majority of the schools have positive effects with credible intervals lying away from 0. Again, we see large standard errors in Schools 11, 13, and 14, in some part due to smaller network size and variability across schools.

In addition to pooling information across schools, another advantage to fitting HLSMs instead of separate LSMs to each school is that we can estimate an overall effect mean and variance for each β, parameterized as μ_i and σ_i^2, respectively, in Eq. (18.7). The posterior means of μ and σ^2 for each covariate effect are given in Table 18.1. We find overall positive effects of similar experience and beliefs, although both credible intervals include zero, which suggests nonsignificant effects. This is unsurprising given the estimates for σ^2, likely because of the small number of networks in the data. The variable for teaching the same grade has an overall mean that is significant and positive, which suggests a strong effect of this variable across all schools. Finally, we find that σ^2 varies for each covariate; we find the smallest amount of variability in the effect of teaching the same grade, whereas the variability in the effects of similar experience and beliefs were both larger.

Thus LSMs and HLSMs are useful for researchers interested in the effects of covariates on network ties. We illustrated that formal position, such as teaching the same grade, is more important for advice seeking than are beliefs about innovation or experience, and other research has suggested similar findings (e.g., Spillane, Kim, & Frank, 2012). Still, for other relationships, such as coteaching, other variables may be more important, and for other contexts, such as student collaboration, the variables could be completely different. HLSMs can also estimate the variability across networks in covariate effects as well as network-level effects. These models can also be used to study the breakdown of teams, identifying which attributes contribute most to dissolution of relationship ties.

18.3.2 Hierarchical Mixed Membership Stochastic Block Models

Hierarchical mixed membership stochastic block models (HMMSBM; Sweet, Thomas, & Junker, 2014), like single-network MMSBMs, are most appropriate for networks with some amount of block structure. All networks need not have block structure, but a nontrivial proportion should have this structure for these models to fit well.

When we first introduced MMSBMs, we discussed a parameter that measures the amount of insularity that exists among subgroups. HMMSBMs allow us to estimate this measure across a sample of networks so that networks can be

Table 18.1 Posterior means and 95% confidence intervals for overall mean and variance (μ and σ^2) for each covariate effect β

Covariate β	Overall mean μ	Overall variance σ^2
SimExp	0.25 [−0.73, 1.08]	3.97 [2.10, 7.50]
SimBeliefs	0.15 [−0.90, 1.05]	4.27 [2.12, 8.25]
SameGrade	1.24 [1.05, 2.09]	2.41 [1.64, 3.78]

compared with one another. In fact, this measure is a relative measure, so it must be estimated in comparison with other networks.

We can write this model in the same format as Eq. (18.3):

$$P(Y|B,S,R) = \prod k \prod_{ij} P(Y_{ijk}|S_{ijk}, R_{jik}, B_k)P(Y_{ijk} = 1) = S^t_{ijk}B_kR_{jik}; \tag{18.8}$$

however, because our focus is on parameters at a different level, we usually present the HMMSBM using a hierarchical Bayesian framework. Although this may appear unconventional to some readers, models written in this way can be quite intuitive for understanding how data arise from the model. We write a HMMSBM as

$$\begin{aligned} Y_{ijk} &\sim \text{Bernoulli } (S'_{ijk}B_kR_{jik}), \\ S_{ijk} &\sim \text{Multinomial } (\theta_{ik}, 1), \\ R_{jik} &\sim \text{Multinomial } (\theta_{jk}, 1), \\ \theta_{ik} &\sim \text{Dirichlet } (\gamma/g), \\ B_{\ell mk} &\sim \text{Beta } (a_{\ell m}, b_{\ell m}), \\ \gamma &\sim \text{Gamma } (c, d), \end{aligned} \tag{18.9}$$

where B, R, and S have the same definitions as in Eq. (18.8). Notice that, in this model, we assume that the networks all have the same value of γ, but we could also specify a model in which γ varies across networks and c and d would be estimated. The same is true for the values of B; we could let the value of B vary across networks, adding an additional layer of hierarchy.

These models are useful for understanding how individuals in networks self-group. Studying across networks, we could look at how the number of subgroups varies across groups. In addition, we could measure the insularity of subgroups across networks. As mentioned in Sect. 18.2, γ is a measure of subgroup integration; networks with low levels of γ are very insular, and networks with high values of γ are integrated (see Fig. 18.2). We might be interested in how values of γ are similar or different across networks as a result of some external variable.

For example, consider an intervention in which collaboration networks of teachers are randomly assigned to treatment or control conditions. The experiment's aim is to change the way teachers interact. Teachers currently collaborate within departments, and school networks consist of highly insular subgroups. The treatment encourages collaborations across department, so we expect to see more integration across subgroups in treated networks. To illustrate this situation, we use simulated data.

We simulated data from the following HMMSBM:

$$\begin{aligned} Y_{ijk} &\sim \text{Bernoulli}\ (S'_{ijk} B_k R_{jik}), \\ S_{ijk} &\sim \text{Multinomial}\ (\theta_{ik}, 1), \\ R_{jik} &\sim \text{Multinomial}\ (\theta_{jk}, 1), \\ \theta_{ik} &\sim \text{Dirichlet}\ (\gamma/g), \\ \gamma &= 0.1 + 0.5X_k, \end{aligned} \tag{18.10}$$

where X_k is the treatment indicator and B is a 3×3 matrix with diagonal entries equal to 0.6 and off-diagonal entries equal to 0.005 and is the same across networks. This means that teachers are likely to collaborate only with teachers in their same department.

We simulated 30 networks with 20 nodes per network; 15 networks are in each condition. Figure 18.5 shows a sample of 10 of these networks. The networks in top row are in the control condition, and the three subgroups are quite insular; the bottom row shows some of the treated networks whose subgroups are much more integrated.

We fit the following HMMSBM to the simulated data:

$$\begin{aligned} Y_{ijk} &\sim \text{Bernoulli}\ (S'_{ijk} B_k R_{jik}), \\ S_{ijk} &\sim \text{Multinomial}\ (\theta_{ik}, 1), \\ R_{jik} &\sim \text{Multinomial}\ (\theta_{jk}, 1), \\ \theta_{ik} &\sim \text{Dirichlet}\ (\gamma/g), \\ \gamma &= \gamma_0 + \alpha X_k, \\ \gamma_0 &\sim \text{Gamma}\ (1, 10), \\ \alpha &\sim \text{Uniform}\ (0, 1). \end{aligned} \tag{18.11}$$

Note that although γ can be any positive number, network subgroup structure when $\gamma > 1$ is generally integrated, and recovering γ when $\gamma > 1$ is difficult owing to a lack of subgroup structure (Sweet & Zheng, 2017). In fact, when $\gamma = g$, the Dirichlet distribution is equivalent to a multivariate uniform distribution.

As for fitting a HMMSBM, we fit the model using a MCMC algorithm using Gibbs updates when possible, and Metropolis updates otherwise; additional details can be found in Sweet et al. (2014). We note that there is an identifiability issue in the model because the value of γ and entries of B are conflated; integrated subgroups result from both large values of γ and certain entries of B.[1] Sweet and Zheng

[1] Integrated networks are generated when diagonal entries of B are close to off-diagonal entries to B.

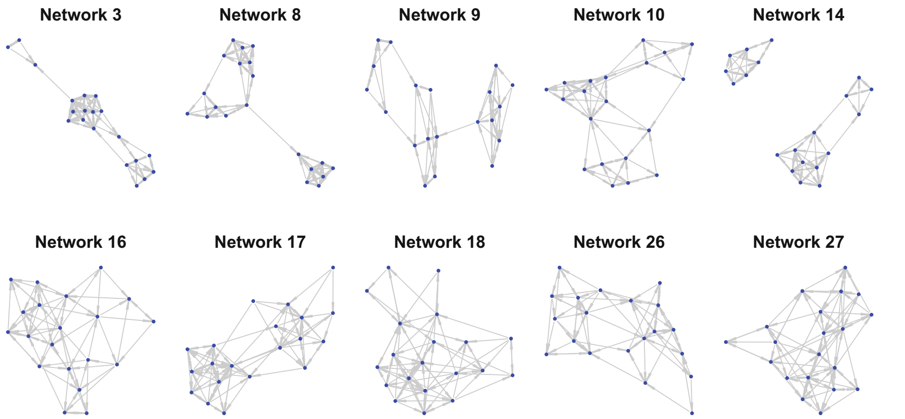


Fig. 18.5 Sample of networks from (*top*) treatment and (*bottom*) control conditions generated from a hierarchical mixed membership stochastic blockmodel

Table 18.2 Hierarchical mixed membership stochastic blockmodel posterior summaries

	γ_0	α
Truth	0.1	0.5
Mean	0.11	0.48
95% CI	[0.09, 0.14]	[0.42, 0.54]

(2017) explored this issue in detail and recommended fixing the B matrix to accurately recovery γ. Despite being an overparameterized model, the HMMSBM does not have any other identifiability issues, and MCMC chains tend to converge quickly. For example, both γ_0 and α converge in the first few hundred steps.

Both γ_0 and α are well recovered. Posterior means and credible intervals are given in Table 18.2. As shown by the generative model of Eq. (18.10) and corroborated here, we find that networks in the control condition have small values of γ and that the positive treatment effect increases the value of γ in the treatment networks to produce more integrated subgroups.

Using HMMSBMs to estimate intervention effects may be important for understanding not only whether certain interventions are effective in changing how individuals interact but also how interventions change network structure. HMMSBMs in particular can address whether network subgroup structure was affected. Similarly, in an observational study, we might be interested in exploring the variability in subgroup insularity across teams.

In addition, how subgroups interact within a larger network can inform team research in other ways. For example, researchers might be interested in how covariates affect network integration. Sweet and Zheng (2016) introduced a model in which $\gamma_k = \exp(\beta' X_k)$ to relate friendship networks of students in classrooms to teacher classroom management styles. Analogous studies could examine larger collaboration networks and the effects of integrated subgroups on group performance or, conversely, the effects of team attributes on subgroup integration.

Of course, the caveat is that these models require networks with some amount of subgroup structure; a sample of networks without any subgroup structure is not appropriate for these models. Focusing on networks with naturally occurring subgroups, such as adolescent friendship networks or networks of individuals who are organized into small communities, such as departments, will help produce network data with the necessary subgroup structure to use these models.

18.4 Conclusion

Social network models allow researchers to estimate relationships between nodes or network features and network ties. Conditionally independent network models, such as latent space models and stochastic blockmodels, model the value of the tie as conditional on some latent structure. This allows researchers to easily investigate the effects of various covariates on tie values and extend these models for use with multiple networks.

Whereas single-network models address research questions about the nodes and ties, hierarchical network models instead focus on generalizing those research questions as well as answering additional questions about variability among networks and attributes about the networks themselves. Network-level research questions include exploring the effects of an intervention on networks or relating some network attribute to various network structures, such as the way in which subgroups cluster.

Regardless of the research question, network models are particularly suited for the complex relationships that exist among teams and groups with their intractable interdependencies. In addition to analyzing network data about collaboration, analyzing network data about other types of relationships, such as interactions, friendship, and even antagonistic relationships, can help researchers investigate and assess teams.

One limitation, however, is that these models are cross-sectional; network ties are not changing. Modeling every interaction, or even a subset of interactions, among a group of individuals poses an even greater methodological challenge. One approach is to treat the series of interactions as a count, using valued-ties perhaps, and then use descriptive methods or a model presented in this or another chapter. Another approach is to use a dynamic network model, such as the relational event model presented in Schector and Contractor (Chap. X) or temporal network models to study interactions over time.

Acknowledgements This research was supported in part by the Institute of Education Sciences (IES) Hierarchical Network Models for Education research grant R305D120004 to Carnegie Mellon University and by Institute of Education Sciences (IES) Hierarchical Network Models: Mediation and Influence grant R305D150045.

References

Adhikari, S., Junker, B. W., Sweet, T. M., & Thomas, A. C. (2014). *HLSM: Hierarchical latent space network model (R package version 3.0.0)*. Pittsburgh, PA, USACarnegie Mellon University.

Airoldi, E., Blei, D., Fienberg, S., & Xing, E. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research, 9,* 1981–2014.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). New York, NY, USA: Taylor & Francis

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association, 97,* 1090–1098.

Holland, P., Laskey, K., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks, 5,* 109–137.

Hopkins, M., Lowenhaupt, R., & Sweet, T. M. (2015). Organizing instruction in new immigrant destinations: District infrastructure and subject-specific school practice. *American Educational Research Journal, 52*, 408–439. doi:10.3102/0002831215584780

Kolaczyk, E. (2009). *Statistical analysis of network data: Methods and models*. New York, NY, USA: Springer.

Pitts, V., & Spillane, J. (2009). Using social network methods to study school leadership. *International Journal of Research & Method in Education, 32,* 185–207.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA, USA: Sage.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Snijders, T. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology, 21,* 149–172.

Spillane, J. P., Kim, C. M., & Frank, K. A. (2012). Instructional advice and information providing and receiving behavior in elementary schools exploring tie formation as a building block in social capital development. *American Educational Research Journal, 119,* 72–102.

Sweet, T. M., Thomas, A. C., & Junker, B. W. (2013). Hierarchical network models for education research: Hierarchical latent space models. *Journal of Educational and Behavioral Statistics, 38,* 295–318.

Sweet, T. M., Thomas, A. C., & Junker, B. W. (2014). Hierarchical mixed membership stochastic blockmodels for multiple networks and experimental interventions. In E. Airoldi, D. Blei, E. Erosheva, & S. Fienberg (Eds.), *Handbook on mixed membership models and their applications* (pp. 463–488). Boca Raton, FL, USA: Chapman & Hall/CRC.

Sweet, T. M., & Zheng, Q. (2015). Multilevel social network models: Incorporating network-level covariates into hierarchical latent space models. In J. Harring, L. Stapleton, & S. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 361–389). Charlotte, NC, USA: Information Age.

Sweet, T. M., & Zheng, Q. (2016). The hierarchical mixed membership stochastic blockmodel with network level covariates. Manuscript revision under review.

Sweet, T., & Zheng, Q. (2017). A mixed membership model-based measure for subgroup integration in social networks. *Social Networks*, 48, 169–180.

Wang, P., Robins, G., Pattison, P., & Lazega, E. (2013). Exponential random graph models for multilevel networks. *Social Networks, 35,* 96–115.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). New York, NY, USA: Cambridge University Press.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika, 61*, 401–425. doi:10.1007/BF02294547

Chapter 19
Network Models for Teams with Overlapping Membership

Mengxiao Zhu and Yoav Bergner

Abstract Systems of teams with overlapping members arise in employment, training, and educational contexts. Team interdependence in these systems can confound analyses that aim to account for both individual and team attributes in studying team formation and performance. This chapter introduces bipartite networks for modeling teams with overlapping members. In these networks, individuals and teams are represented by two different types of nodes with links representing team affiliation. Two methods for analysis of bipartite networks with individual and team attributes are reviewed, exponential random graph models (ERGMs) and correspondence analysis (CA). Examples, discussions, and comparisons are provided for both methods.

Keywords Teams · Network model · Bipartite network · Exponential random graph models (ERGMs) · Correspondence analysis

19.1 Introduction

A team is a set of individuals working collaboratively towards a goal. In employment, training, and educational contexts, individuals may participate in multiple teams, at the same time or sequentially. In such systems of teams with overlapping members, a number of interesting questions naturally arise. For example, when team assembly is voluntary, what factors drive this process? Are individuals more likely to join with teammates who share common attributes, such

This work was conducted while Yoav Bergner was employed with Educational Testing Service.

M. Zhu (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: mzhu@ets.org

Y. Bergner
New York University, New York, NY, USA
e-mail: yoav.bergner@nyu.edu

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_19

as age or gender? To what extent do social connections or prior collaboration history impact teammate selection? For either self-assembled or assigned teams, one may also be interested in using individual attributes, connections, and history to predict performance. Team-level attributes and history may also inform performance.

The statistical dependencies associated with overlapping membership in systems of teams can substantially complicate the analysis of teaming and performance (Kozlowski & Klein, 2000). In the past, researchers have sometimes steered around this problem by ignoring the overlap between teams or by choosing only teams without overlapping members (Cummings & Cross, 2003; Oh et al., 2004). However, such approaches neglect valuable information and ignore a realistic feature of many organizations.

Network models, specifically bipartite graphs, can be used effectively to model teams with overlapping members. In network data, the observations are nodes and links, as well as node or link attributes. In a bipartite network (Wasserman & Faust, 1994), there are two different types of nodes, and links run only between the nodes of different types. It is thus quite natural to represent individuals and teams as the two different types of node, with links representing team membership.

In the remainder of this chapter, we introduce the use of bipartite networks to represent systems of teams with overlapping members and discusses two statistical analysis techniques that are appropriate for such network data. Specifically, we describe the application of exponential random graph models (ERGM, also known as p^*)—as extended to bipartite networks (Robins, Pattison, Kalish, & Lusher, 2007; Wang, Sharpe, Robins, & Pattison, 2009)—and correspondence analysis (Nenadic & Greenacre, 2007; Wasserman, Faust, & Galaskiewicz, 1990). We also compare the characteristics of these two methods and their applications in studying team formation and performance.

19.2 Bipartite Network Models for Teams

Consider the system of teams in Fig. 19.1a. This system can be represented using the bipartite network shown in Fig. 19.1b, where solid circles indicate individuals and triangles indicate teams. The links in the bipartite network represent the team membership. Person a is a member of Team 1 and also of Team 3.

A bipartite network B of individuals and teams can be represented using an affiliation matrix $\mathcal{A}$. By convention, each row of the matrix represents an individual, and each column represents a team. The values a_{ij} in this affiliation matrix indicate the team membership,

$$a_{ij} = \begin{cases} 1 & \text{if person } i \text{ is a member of team } j \\ 0 & \text{otherwise} \end{cases}.$$

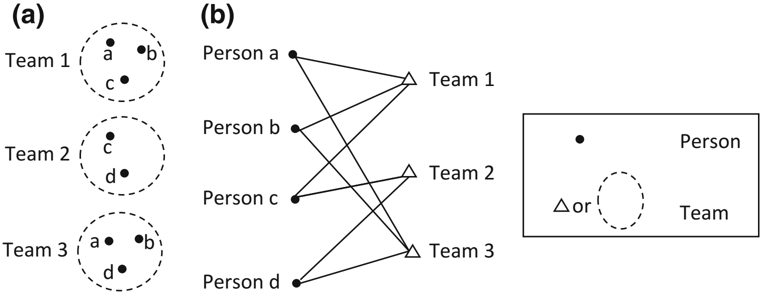
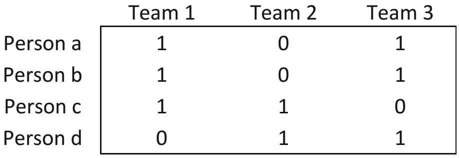


Fig. 19.1 Bipartite network model fort teams with overlapping members

Fig. 19.2 Affiliation matrix representation for bipartite networks



	Team 1	Team 2	Team 3
Person a	1	0	1
Person b	1	0	1
Person c	1	1	0
Person d	0	1	1

The bipartite network of Fig. 19.1 can be represented using the 4 × 3 affiliation matrix in Fig. 19.2. Note that the row marginals $\{a_{i+}\}$ are the number of teams that each individual participated in, and the column marginals $\{a_{+j}\}$ are the team sizes.

Bipartite networks compactly encode both individual and team level attributes while preserving the team membership and related dependency. This is especially important when individuals participate in more than one team. In this case, neither teams nor individuals can be considered independent observations.

Figure 19.3 shows a bipartite network of 159 teams assembled by 168 individuals from a massively multiplayer online role-playing game (MMORPG) (Zhu, Huang, & Contractor, 2013). Black circles indicate players and white triangles indicate teams. In the game, the players self-assembled into teams for combat in the virtual world, and most players joined more than one team. A large group of players, seen in the blob to the right, are connected through shared team membership. However, the smaller components in the left crescent show that many players form and reform teams from within the same small subgroup.

Analysis of bipartite networks can help identify patterns of team participation and performance. Which individuals tend to work together? What individual or team attributes are related to high performance? In the following sections, we review two methods that can be used to account for dependencies embedded in these network data.

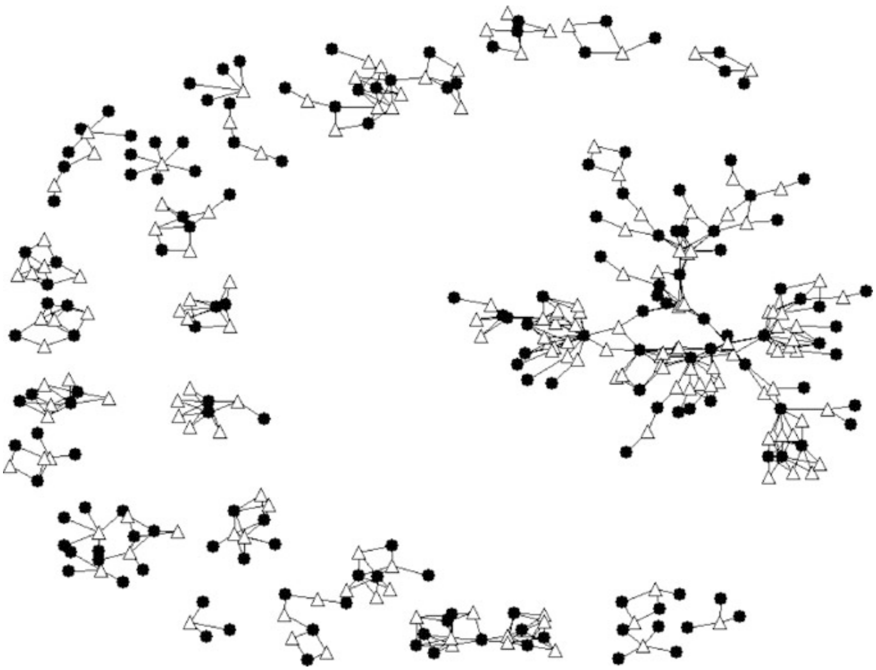


Fig. 19.3 Example of a team network

19.3 ERGM for Bipartite Network

Exponential random graph models (ERGMs), also known as p^* models, can be used to examine and test relatively complex hypotheses about network structures and interactions of node attributes with network structures. ERGMs are a class of stochastic models that share the following general form (Wasserman & Pattison, 1996),

$$P(Y = y) = \frac{1}{k(\theta)} \exp\left(\theta^T g(y)\right),$$

where Y is the network realization, a random variable. Its state space is the collection of all possible networks with the same number of nodes as the observed network, denoted y, and $P(Y = y)$ is the probability of observing y. The vector of network statistics, $g(y)$, can include network structures, such as links, triads, and stars. This vector can also include interactions between node attributes and network structures, such as counts of edges between actors of the same gender. Other relevant vectors are θ, a vector of coefficients, and $k(\theta)$, a normalizing constant calculated by summing $\exp\left(\theta^T g(y)\right)$ over the space of possible networks.

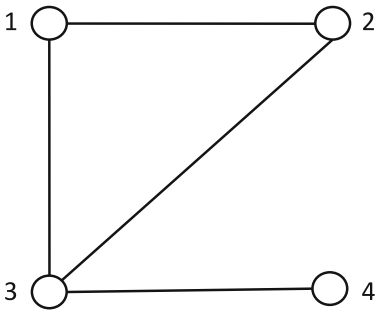


Fig. 19.4 A simple network with four nodes

Consider the network shown in Fig. 19.4. This is an undirected network with four nodes and no self-loops. If we restrict our attention to edges, ignoring other network structures, the vector of network statistics may be written as

$$g(y) = [y_{12}, y_{13}, y_{14}, y_{23}, y_{24}, y_{34}]^T = [1, 1, 0, 1, 0, 1]^T,$$

where each y_{ij} is an indicator for an edge in the network between node i and node j.

For each element in $g(y)$, there is a corresponding coefficient in θ,

$$\theta = [\theta_{12}, \theta_{13}, \theta_{14}, \theta_{23}, \theta_{24}, \theta_{34}]^T$$

The normalizing constant $k(\theta)$ is calculated by adding up $\exp(\theta_{12}y_{12} + \theta_{13}y_{13} + \theta_{14}y_{14} + \theta_{23}y_{23} + \theta_{24}y_{24} + \theta_{34}y_{34})$ over all possible networks with four nodes. There are six possible edges in this undirected network, so there are $2^6 = 64$ possible networks. Therefore, the probability of observing the network in Fig. 10.1 can be expressed as follows:

$$\begin{aligned} P(Y = y) &= \frac{1}{k(\theta)} \exp(\theta_{12}y_{12} + \theta_{13}y_{13} + \theta_{14}y_{14} + \theta_{23}y_{23} + \theta_{24}y_{24} + \theta_{34}y_{34}) \\ &= \frac{1}{k(\theta)} \exp(\theta_{12} + \theta_{13} + \theta_{23} + \theta_{34}). \end{aligned}$$

Specific assumptions about homogeneity and dependence may be imposed on the general ERGM. For instance, under the assumption that all edges in the network are homogenous and independent, one obtains the Bernoulli or Erdős–Rényi model (Erdős & Rényi, 1959),

$$P(Y = y) = \frac{1}{k(\theta)} \exp\left(\theta \sum\nolimits_{ij} y_{ij}\right).$$

In this case, the only network statistic is the count of the edges.

Examples of other dependency assumptions include the $p1$ model (Holland & Leinhardt, 1981), which assumes that dyads, instead of edges, are independent of each other. For undirected networks, the $p1$ model is equivalent to the Bernoulli model. For directed networks, however, the $p1$ model includes both edge statistics and reciprocity statistics. The $p2$ model (Lazega & van Duijn, 1997; Van Duijn, Snijders, & Zijlstra, 2004) extends the $p1$ model by conditioning edges on nodal-level attributes, which is appropriate when nodal attributes are a major driving force of network structures. In both $p1$ and $p2$ models, the homogeneity over the network is assumed, that is, nodes and edges are not distinguished by their indices. Markov dependence (Frank & Strauss, 1986), on the other hand, assumes that the possible edge between node i and j is dependent on any other edges involving either node i or node j.

The extension of ERGMs to bipartite networks (e.g., Agneessens & Roose, 2008; Agneessens, Roose, & Waege, 2004; Faust, Willert, Rowlee, & Skvoretz, 2002; Wang et al., 2009) was particularly useful for analyzing systems of teams with overlapping members.

To illustrate this application, we use the example network in Fig. 19.3. Figure 19.5 shows three examples of network structures and related hypotheses that can be tested for the bipartite dataset (more details available in Zhu et al., 2013).

As before, circles are individuals and triangles are teams. Shaded shapes are now used to indicate when nodal attributes are relevant to the hypothesis. When nodal attributes are included in addition to topological network structures, we refer to the combined object as a network configuration. The plus/minus signs indicate whether the related network configuration is expected to be observed more/less often than by random chance. Finally, the sizes of the nodes indicate high or low values on the attributes. For instance, Fig. 19.5a shows a network configuration that can be used to test a hypothesis involving an individual attribute (skill) and a single teaming relation, such as *H1: High-skilled individuals are less likely to assemble into teams than low-skilled individuals.* The shaded circle represents individuals with the attributes of skill levels and the bigger size of the individual node indicates higher

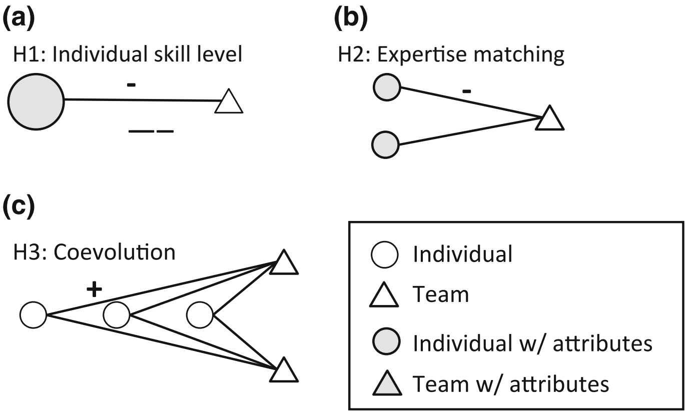


Fig. 19.5 Network configurations and related hypotheses

level on individual skill levels. The minus sign indicates that the structure of higher level individuals joining teams is less likely to be observed. The related network statistic is the *weighted* sum of individual-team links, where the individual's skill is used as the weight.

The network structure in Fig. 19.5b includes two individual nodes with attributes, one team node and teaming links. It can be used to represent the hypothesis, *H2: Individuals are less likely to assemble into teams with others who possess the same expertise.* The related network statistic is the count of the number of structures with two individuals with the same expertise connected to the same team. The minus sign for the network configuration indicates that this network configuration is expected to be observed less frequently than by random chance. Finally, Fig. 19.5c represents *H3: Individuals are more likely to join those with whom they previously collaborated.* The corresponding network structure is two or more individuals joining two teams together.

Hypothesis tests using ERGMs for bipartite networks can be carried out with freely available software packages such as BPNet (Wang et al., 2009). The parameters of Markov ERGMs are estimated using Monte Carlo maximum likelihood techniques (Robins et al., 2007) and are interpreted in a similar way as parameters in ordinary regression models (Wang et al., 2009).

19.4 Correspondence Analysis for Bipartite Networks

Correspondence analysis (CA; Greenacre, 2007) and its extension, multiple correspondence analysis (MCA), were originally developed as multivariate statistical analysis techniques for categorical data. Both are data analysis and visualization tools that take contingency tables as the input and represent the data in a two-dimensional graph. Multiple correspondence analysis can be applied to more than two categorical variables. Network researchers have developed methods to apply correspondence analysis to regular and bipartite network data (e.g., Wasserman et al., 1990; D'Esposito, De Stefano, & Ragozini, 2014; Roberts, 2000; Zhu, Kuskova, Wasserman, & Contractor, 2015).

To apply CA to bipartite networks, the affiliation matrix $\mathcal{A}$, described in Sect. 10.2, becomes the input. Rows represent persons, columns represent teams, and the matrix elements are indicators $\{0, 1\}$ of team affiliation. CA generates a set of scores for rows and columns, which can be used to embed person and team nodes in a lower dimensional space (typically two dimensions). This reduction and visualization of the original dataset enables researchers to cluster similar individuals or teams and to find patterns in the team relations.

The procedure is as follows: given a $p \times t$ affiliation matrix $\mathcal{A}$, generate a set of scores and parameters on the dimensions that are equal to or less than $W = \min(p - 1, t - 1)$. This include a set of p row scores $\{u_{ik}\}$, for $i = 1, \ldots, p$ and $k = 1, \ldots, W$, on each of W dimensions for persons; a set of t column scores $\{v_{jk}\}$, for $i = 1, \ldots, t$ and $k = 1, \ldots, W$, on each of W dimensions for teams; and a set of W principal inertias (eigenvalues) $\{\eta_k^2\}$, for $k = 1, \ldots, W$, presenting the correlation between the rows and columns (Wasserman & Faust, 1994). The scores and principal inertias satisfy the simultaneous equations:

$$\eta_k u_{ik} = \sum_{j=1}^{h} \frac{a_{ij}}{a_{i+}} v_{jk},$$

$$\eta_k v_{jk} = \sum_{i=1}^{g} \frac{a_{ij}}{a_{+j}} u_{ik},$$

where a_{ij} is an element in the affiliation matrix $\mathcal{A}$, and a_{i+} and a_{+j} are the row and column marginals. The principal coordinates (eigenvectors) can be rescaled to get standard coordinates $\tilde{u}$ and $\tilde{v}$ (Greenacre, 1984), with weighted mean of 0 and weighted variance of 1:

$$\tilde{u}_{ik} = u_{ik}/\eta_k,$$

$$\tilde{v}_{jk} = v_{jk}/\eta_k.$$

Two dimensional plots are usually used to display the results from the correspondence analysis (Nenadic & Greenacre, 2007; Wasserman & Faust, 1989). Open source software such as R package **ca** (Nenadic & Greenacre, 2007) and commercial software programs, such as SAS, have functions for correspondence analysis and multiple correspondence analysis.

For bipartite networks, the data points in the plots represent nodes in the networks, and the point locations are the standardized principal coordinates. An example of the correspondence analysis results for a network similar to Fig. 19.3 is shown in Fig. 19.6, which was generated using R packages **ca** and **ggplot2** (Wickham, 2012). The axis labels "First Eigenvector" and "Second Eigenvector" indicate the first two dimensions of the standardized coordinates. Circles represent individuals and triangles represent teams, while the numbers near the points are identifiers.

For nodes of the same type, that is, between individuals or between teams, proximity implies similarity. Proximity between an individual P and a team T obtains if similar individuals are members of T or if similar teams include P or both. The clusters in the correspondence analysis plot are consistent with the graph components in the network shown in Fig. 19.3; there is one big cluster and several small clusters. Correspondence analysis provides a useful way to graphically represent bipartite networks and the similarities among different types of nodes in the low-dimensional space (D'Esposito et al., 2014).

The power of correspondence analysis goes beyond the visual analysis of bipartite networks. Multiple correspondence analysis (Nenadic & Greenacre, 2007; Wasserman et al., 1990) can be used to add individual and team-level attributes. A multiple indicator matrix is constructed as follows. For each nonzero cell in an

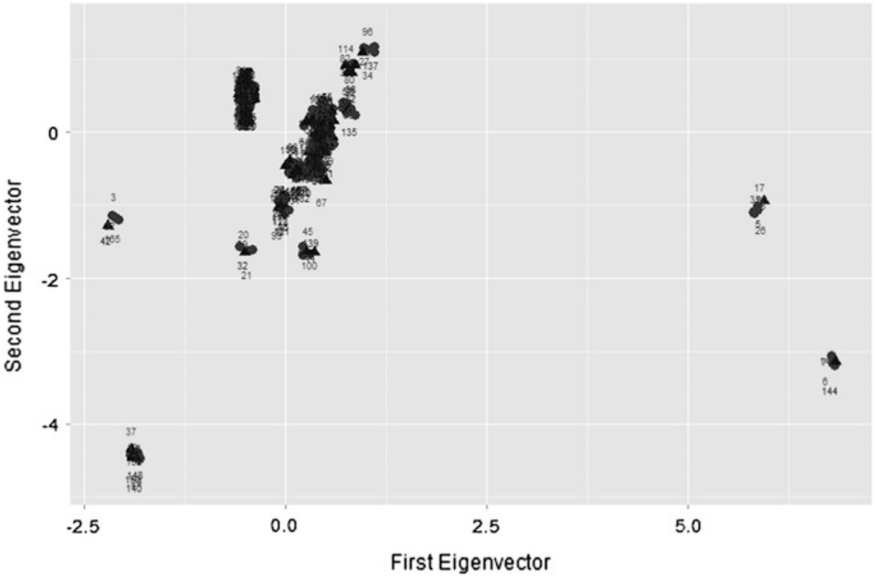


Fig. 19.6 Correspondence analysis for the bipartite network data

affiliation matrix, that is, each individual-team membership relation, a row is created in the multiple indicator matrix. Individual and team-level attributes must be categorized, if they are not already categorical, and dummy coded as a vector of indicator variables. The full set of attribute indicator variables comprise the columns of the multiple indicator matrix. For instance, if an individual gender attribute takes the values female or male, two corresponding columns will be created. The constructed multiple indicator matrix is the input for MCA.

For the data in Fig. 19.3, we conducted a multiple correspondence analysis on one individual attribute, gender, and one team attribute, team performance. Individual gender corresponded to the real-life gender of the game players (as opposed to their online avatars). Team performance was measured by the number of monsters killed by the team in the virtual games, and categorized as low, medium, and high performance. Analysis was performed using the *mca* routine in R package **ca** (Nenadic & Greenacre, 2007). The principal inertias indicate that there are three dimensions in this example. The first two explain 36.3 and 33.3%, respectively, of the observed variance.

Individual and team-level node attributes are plotted in the reduced two-dimensional space in Fig. 19.7. The circles represent gender, and the triangles represent levels of team performance. Proximity of points indicates stronger associations between levels of different attributes. Looking holistically, we see that male players tend to be affiliated with teams with low or medium performance teams and female players with high-performing teams. In fact, all of this information arises from variance associated with the first dimension (x-axis) in the MCA. When

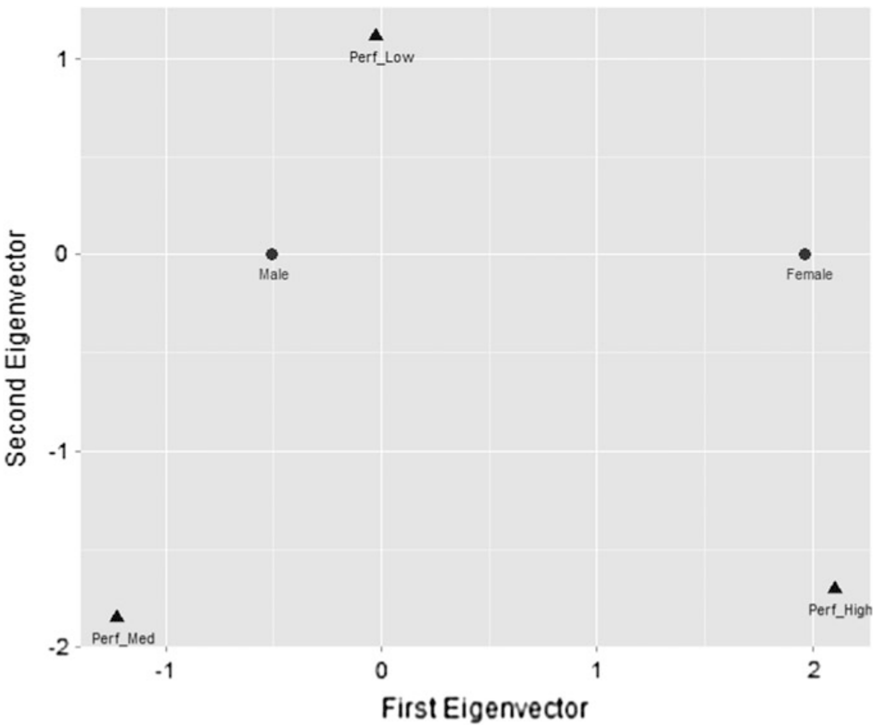


Fig. 19.7 Multiple correspondence analysis of gender and team performance

projected onto the y-axis, the results do not differentiate between genders but do indicate similarity between medium and high performance team attributes.

19.5 Discussion and Conclusion

In this chapter, we introduced bipartite networks for modeling teams with overlapping members. Because of their shared membership, these teams cannot be considered independent of each other. Analysis of team formation patterns and relationships between individual and team attributes necessitates a modeling framework that can properly account for this interdependence. We reviewed two applicable methods, exponential random graph models for bipartite networks, and correspondence analysis. We now turn to a brief discussion comparing the two methods.

ERGMs for bipartite networks use network structures (e.g., edges) and structure-attribute configurations (e.g., edges between individuals with matching attributes) as the basic building block of analysis. The major strength of this method is its ability to identify statistically significant configurations, that is, those that occur more frequently than expected by chance alone. Thus, this method is often used to answer research questions related to what types of collaborations are observed in the current system, what kinds of individuals tend to collaborate with each other, or what kinds of teams tend to attract more members. However, due to present limitations in these models, it is not possible to test for interactions between individual and team attributes. Furthermore, estimation of ERGMs is computationally costly, especially for large networks.

In contrast, correspondence analysis can flexibly account for attributes at the individual or team level as well as interactions between levels of each type. The results may be explored using two-dimensional plots rather than tabulated numbers. For example, CA can be used to probe research questions such as, what types of teams are high-performing, or what types of individuals tend to join high-performing teams. Correspondence analysis is essentially exploratory and does not provide significance tests for the observed associations. Moreover, it is limited to categorical data, which poses potential problems when used with continuous attributes.

Given their advantages and disadvantages, these two methods may best be used to answer different research questions or in a complementary fashion. Correspondence analysis can be easily applied to big datasets. The data exploration function and relatively low-cost estimation process for correspondence analysis make it a powerful tool to explore the data and discover relations among variables. This kind of exploration may provide some foundations for more costly ERGM building and estimation. When well-founded hypotheses can be expressed as network configurations, ERGM analysis can provide more robust tests of significance.

References

Agneessens, F., & Roose, H. (2008). Local structural patterns and attribute characteristics in 2-mode networks: *p** models to map choices of theater events. *Journal of Mathematical Sociology, 32,* 204–237.

Agneessens, F., Roose, H., & Waege, H. (2004). Choices of theatre events: *p** models for affiliation networks with attributes. *Developments in Social Network Analysis, Metodoloski Zvezki, 1,* 419–439.

Cummings, J. N., & Cross, R. (2003). Structural properties of work groups and their consequences for performance. *Social Networks, 25*(3), 197–210.

D'Esposito, M. R., De Stefano, D., & Ragozini, G. (2014). On the use of multiple correspondence analysis to visually explore affiliation networks. *Social Networks, 38*, 28–40. doi:10.1016/j.socnet.2014.01.003

Erdős, P., & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae, 6,* 290–297.

Faust, K., Willert, K. E., Rowlee, D. D., & Skvoretz, J. (2002). Scaling and statistical models for affiliation networks: Patterns of participation among Soviet politicians during the Brezhnev era. *Social Networks, 24,* 231–259.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association, 81*(395), 832–842.

Greenacre, M. (1984). *Theory and applications of correspondence analysis*. London, UK: Academic Press.

Greenacre, M. (2007). *Correspondence analysis in practice* (2nd ed.). New York, NY: CRC Press.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association, 76,* 33–65.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco, CA: Jossey-Bass.

Lazega, E., & van Duijn, M. (1997). Position in formal structure, personal characteristics and choices of advisors in a law firm: A logistic regression model for dyadic network data. *Social Networks, 19,* 375–397.

Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software, 20*(3). Retrieved October 20, 2016 from http://www.jstatsoft.org/

Oh, H., Chung, M.-H., & Labianca, G. (2004). Group social capital and group effectiveness: The role of informal socializing ties. *Academy of Management Journal, 47*(6), 860–875.

Roberts, J. M. (2000). Correspondence analysis of two-mode network data. *Social Networks, 22,* 65–72.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (*p**) models for social networks. *Social Networks, 29*(2), 173–191.

Van Duijn, M. A. J., Snijders, T. A. B., & Zijlstra, B. J. H. (2004). p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica, 58,* 234–254.

Wang, P., Sharpe, K., Robins, G. L., & Pattison, P. E. (2009). Exponential random graph (*p**) models for affiliation networks. *Social Networks, 31*(1), 12–25.

Wasserman, S., & Faust, K. (1989). Canonical analysis of the composition and structure of social networks. *Sociological Methodology, 19,* 1–42.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, England: Cambridge University Press.

Wasserman, S., Faust, K., & Galaskiewicz, J. (1990). Correspondence and canonical analysis of relational data. *Journal of Mathematical Sociology, 1,* 11–64.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and *p**. *Psychometrika, 61*(3), 401–425.

Wickham, H. (2012). ggplot2 [Computer software]. Retrieved October 20, 2016, from https://cran.r-project.org/

Zhu, M., Huang, Y., & Contractor, N. S. (2013). Motivations for self-assembling into project teams. *Social Networks, 35*(2), 251–264. doi:10.1016/j.socnet.2013.03.001

Zhu, M., Kuskova, V., Wasserman, S., & Contractor, N. (2015). Correspondence analysis of multirelational multilevel network affiliations: Analysis and examples. In E. Lazega & T. A. B. Snijders (Eds.), *Multilevel network analysis for the social sciences—Theory, methods and applications*. New York, NY: Springer.

Chapter 20
Linking Team Neurodynamic Organizations with Observational Ratings of Team Performance

Ronald Stevens, Trysha Galloway, Jerry Lamb, Ron Steed, and Cynthia Lamb

Abstract We have investigated the correlations between the levels of team resilience as determined by expert raters and the degree of the teams' neurodynamic organization determined by electroencephalography (EEG). Neurophysiologic models were created from submarine navigation teams that captured their dynamic responses to changing task environments during required simulation training. The teams were simultaneously rated for resilience by two expert observers using a team process rubric developed and adopted by the U.S. Navy. Symbolic neurodynamic representations of the power levels in the 1–40 Hz EEG frequency bands were created each second from each crew member. These symbols captured the EEG power of each team member in the context of the other team members and also in the context of the task. Quantitative estimates of the changes in the symbol distributions over time were constructed by a moving window of Shannon entropy. Periods of decreased entropy were observed when the distribution of symbols in this window became smaller, for example, when there were prolonged and restricted relationships between the EEG power levels among the crew members, that is, less neurodynamic flexibility. Team resilience was correlated with the neurodynamic entropy levels. The correlation sign, however, depended on the

R. Stevens
UCLA School of Medicine, Los Angeles, CA, USA
e-mail: ron@teamneurodynamics.com

R. Stevens · T. Galloway (✉)
The Learning Chameleon Inc, Culver City, CA, USA
e-mail: trysha@teamneurodynamics.com

J. Lamb
Naval Submarine Medical Research Laboratory, Groton, CT, USA
e-mail: jerry.c.lamb.civ@mail.mil

R. Steed
UpScope Consulting Group, Mysti, CT, USA
e-mail: ronaldsteed@upscopeconsulting.com

C. Lamb
URS Federal Technical Services Inc, Philadelphia, PA, USA
e-mail: clamb@egginc.com

A.A. von Davier et al. (eds.), *Innovative Assessment of Collaboration*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-33261-1_20

training segment with negative correlations during the presimulation briefing and positive correlations in the scenario training segment. These studies indicate that neurodynamic representations of teams can be generated that bridge the microscales of EEG measurement with macroscales of behavioral ratings. From a training perspective, the results suggest that neurodynamic rigidity (i.e., everybody on the same page) might be beneficial while teams are preparing for the simulation, but during the scenario, increased neurodynamic flexibility contributes more to team resilience.

Keywords Team neurodynamics · EEG · Team resilience · Synchrony · Shannon entropy · Social dynamics · Symbolic modeling

20.1 Introduction

Our understanding of how to assemble, train, and improve teams' performance has been slowed by a lack of quantitative and objective measures of teamwork. Currently most evaluations of teams performing natural tasks rely on experts who observe and rate teams across important but quantitatively vague dimensions like leadership, team structure, and situation monitoring using vetted rubrics. One widely used evaluation rubric is the Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) program, which was developed for evaluating teams in the health care domain (Baker, Amodeo, Krokos, Slonim, & Herrera, 2009). A second rubric in the military domain is the Submarine Team Behavior Toolkit (STBT), which became available when the Naval Submarine Medical Research Laboratory began an extensive effort to provide the submarine force with a way to improve operational performance by focusing, not on human error per se, but on human variability, which considers not only the action but also the context within which that action occurred.

Observational/behavioral ratings like TeamSTEPPS and STBT tend to rely on macrofeatures of team performance by summarizing observations over extended periods of time. Although the shorter term dynamics of the team are implicitly acknowledged in the resulting ratings, the dynamical details are often lost. As a result, the momentary dynamics of teams performing in natural situations have been largely unexplored.

Recent technological advances in the physiologic and behavioral monitoring of humans are providing new ways of capturing team performance data over very short timescales and are leading to new conceptualizations of teamwork. For instance, changes in the regular pinging of a heart rate monitor may simultaneously trigger similar brain activities in the visual, auditory, and cortical regions of the brains of all team members, that is, a form of natural synchronization. Such synchronization has been repeatedly seen with subjects viewing movie clips (Hasson, Nir, Levy, Fuhrmann, & Malach, 2004), especially when those clips contained emotionally rich scenes (Dmochowski, Sajda, Dias, & Parra, 2012; Nummenmaa et al., 2012). The naturalistic setting of the stimuli in these studies suggested that these ideas of group synchrony might be applicable to teams performing complex tasks in natural settings with team members exhibiting neurodynamic entrainments to particularly important segments of the task.

Teams differ from individuals viewing a movie in important ways, however. Although task signals may simultaneously arrive to each member of the team, the signal information may be perceived differently depending on each member's experiences and responsibilities within the team. Teams can also actively shape the story line as each team member influences, and is influenced by, the others through social coordination. These social coordination activities lead to the generation of a second set of information signals, not from the task or the environment but from other team members while they try to understand each other.

This understanding is derived from the information exchanged between members and is packaged in words (Cooke, Gorman, & Kiekel, 2008) and nonverbal social interactions (Menoret et al., 2014) like gestures (Schippers, Roebroeck, Renken, Nanetti, & Keysers, 2010), posture (Shockley, Santana, & Fowler, 2003), facial expressions (Anders, Heinzle, Weiskopf, Ethofer, & Haynes, 2011), and even periods of silence, all of which contribute to the overall team dynamics.

It is not surprising that neurophysiologic processes are the underpinnings of the coordination dynamics seen in teams, for instance, speaker–listener couplings (Stephens, Silbert, & Hasson, 2010). Multiple neuromarkers of social coordination have also been described in the 9–12 Hz (or alpha) frequency range (Tognoli & Kelso, 2013), which include the 10.9 Hz phi complex, which is modulated by intentional coordination (Tognoli, Lagarde, De Guzman, & Kelso, 2007), and the medial left and right mu EEG components in the alpha (9–11 Hz) and beta (15–20 Hz) frequencies, which may represent activities associated with the human mirror neuron system (Oberman, Pineda, & Ramachandran, 2007; Pineda, 2008). The mirror neuron system is a collection of neurons that respond to actions we see in others. These neurons are active both when a person executes a motor act and when he or she observes another individual performing that act (Caetano, Jousmaki, & Hari, 2007; Rizzolatti, Fogassi, & Gallese, 2001). Through this system, the changing sequence of actions by one person leads to sequences of actions in others—a form of social "resonance" (Schippers et al., 2010).

Although these and similar studies reveal the low-level details of social coordination, the impact of these studies on guiding the process and evaluation of teamwork has been minimal. One reason is that these microlevel speech, gesture, posture, and neurodynamic variables are short-lived and show weak domain or task specificity and cannot be easily linked to the macrolevel observations of raters, the gold standard of team evaluation.

An approach for extending the usefulness of these short-lived activities for measuring team performance would be to view them as hierarchies of fast and slow variables (Flack, 2012). Slow variables, as the name suggests, arise from mechanisms that naturally integrate over faster microscopic dynamics and represent some average of the noisier activities below. For instance, as neurodynamic hierarchies are transited upward from faster scales to slower scales, what would be lost in the mechanistic details of neuronal spike generation and propagation would be gained by tighter relationships with more easily recognized, observer-defined variables such as team coherence, flexibility, or resilience. In this way, the more "intermediate-level" representations could provide a meaningful bridge between the millisecond scales of human brain processing and the observational performance estimates of expert facilitators. *Intermediate representation* is a term borrowed from computer programming that describes a language partway between the source and target languages. A good intermediate representation is one that is fairly independent of the source and target languages so that it maximizes its ability to be repeatedly used in different situations.

Our hypothesis has been that meaningful intermediate representations might be developed spanning timescales of seconds to minutes that would bridge the fast dynamics of common neurophysiologic markers of social coordination with the slower performance variables that arise from behavioral observations like STBT and TeamSTEPPS. These models could begin to link theory and practice in an understandable way, would be applicable to many different team settings, and might serve as objective measures of teamwork.

Several years ago, we explored an information/organization-centric approach for quantitatively mapping the neurophysiologic organizations of teams as a way of relating their fluctuating dynamics to team activities, communications, and performance (Stevens, Galloway, Wang, & Berka, 2011; Stevens & Galloway, 2014, 2015). The goal was to develop data streams that had internal structure(s) with temporal information about the present and past organization, function, and performance of the teams and members of the teams.

Electroencephalography (EEG) was chosen for these studies as it provides real-time and high-resolution temporal measures in an unobtrusive fashion. EEG is the recording of the brain's electrical activity at different regions along the scalp. The rhythmic patterns in the electrical oscillations from different brain regions contain signals representing complex facets of brain activity, many of which reside in the 1–200 Hz frequency range (Buzaki, 2006). Commonly described frequency bands include (a) delta (~ 1–5 Hz), often associated with deep sleep and perhaps with a role in the inhibition of sensory stimuli interfering with internal concentration (Harmony, 2013); (b) theta (~ 7 Hz), related to the processing of episodic information, predictive navigation, and memory encoding and retrieval (Battaglia, Sutherland, & McNaughton, 2004; O'Keefe & Dostrovsky, 1971); (c) alpha (~ 10 Hz), the dominant EEG frequency in the awake human brain, and though primarily thought of as a marker of visual attention, its significance has expanded to one of attention in general, and perhaps prioritizing visual stimuli (Palva & Palva, 2007; Bonnefond & Jensen, 2015); (d) beta (~ 20 Hz), reflecting the cognitive control of motor processes and perhaps top-down brain processes in general; and (e) gamma (>30 Hz), involved in attention, memory encoding, and retrieval and which may operate by transmitting temporal sequences of information across brain regions—gamma oscillations are often nested or phase-locked to theta and/or alpha rhythms (Lisman & Jensen, 2013).

Our approach for modeling such dynamics was to create a symbol each second that showed each team member's power levels at each 1 Hz EEG frequency at different sites on the scalp. A sequence of these symbolic representations of EEG power that spanned the length of the performance would then contain a second-by-second neurodynamic history of the team, the resolution of which would depend on the number of frequencies and channels analyzed.

To the extent that the task activities and team member interactions are predictable, the structure of this symbol stream might be relatively smooth. More interesting segments in these data streams, perhaps those with the most structure, might arise in response to acute or chronic changes to the team/task when the team as a whole became entrained either by the task or other team members' interactions. Under these conditions, a more limited set of neurodynamic relationships might develop among team members and persist. The questions posed for this study were whether we could detect such across–team member persistent neurodynamic structures and whether the frequency, magnitude, and/or duration of these segments could be linked to expert estimates of team performance.

20.2 Methods

20.2.1 Submarine Piloting and Navigation Simulations

The tasks for these studies were required submarine piloting and navigation (SPAN) simulations where the goal was to safely pilot a submarine into or out of port. Each SPAN session contained three segments. First, there was a briefing ($\sim$20–30 min) where the training goals of the mission were presented along with information on the submarine's position, other ships in the area, weather, the sea state, and the captain's orders for safe operation. The scenario ($\sim$50–120 min) followed and was a more dynamic and evolving task containing easily identified processes of teamwork along with other processes less well defined. One regular task was to periodically establish the ship's position. This process was repeated every 3 min and proceeded through a closing sequence of 1 min to next round, 30 s to next round, standby to mark round, and mark round. The debrief section ($\sim$20–30 min) was an open discussion of what worked, what other options were available, and long- and short-term lessons. The debrief was the most structured training with individual team member reports. This brief–scenario–debrief task structure is not unique to military training but is found in other high-stakes teaming activities like surgery.

20.2.2 Submarine Team Behavioral Toolkit

The research efforts behind the development of STBT originated as a study of submarine mishaps as a way to understand the impacts of emerging complexity on human performance. The research indicated that, in addition to technical skills, deliberate and effective team practices were necessary to manage the wide variety of increasingly complex problems that occur during tactical operations. The result was the development of the STBT, which provides an observational guide for assessing team performance. In developing an overall behavioral rating of team resilience, the STBT observers evaluated teams across a set of five practices that have provided new insights into how submarine tactical teams need to operate at sea. When one or more of these practices were absent, team problem solving suffered in some important way. These practices included dialogue, decision making, critical thinking, bench strength, and problem-solving capacity. Each practice contained multiple behavior threads. For decision making, these were decisiveness and leader detachment, whereas for critical thinking, these were planning and time horizon, setting context, managing complexity, and forceful backup. The presence or absence of these practices was linked to four resilience levels describing how different teams performed in complex environments (Fig. 20.1).

The levels of team resilience (in descending order) were (a) advanced team resilience, where the teams could manage multiple dynamic problems; (b) team-based resilience, where routine activities can be managed even during stress; (c) leader-dependent battle rhythm, where the teams retain their rhythm even under stress, but only because someone takes charge; and (d) unstressed battle rhythm, where teams exhibit a rhythm, but only in the absence of disruptions. Evaluator rankings were made on a scale from 0 to 4.

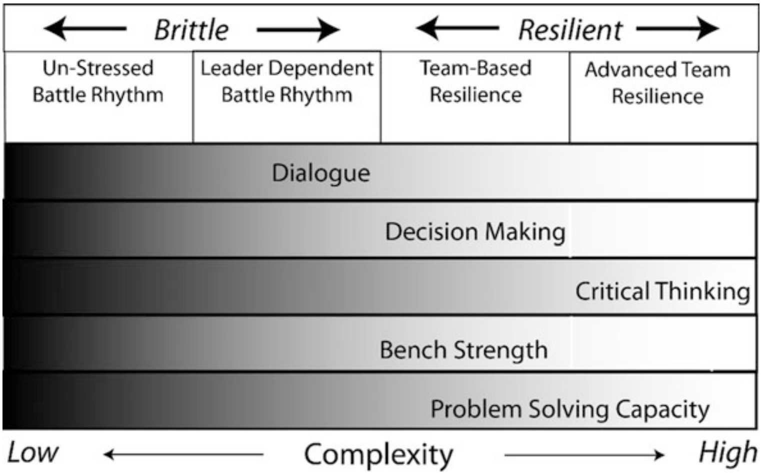


Fig. 20.1 Overview of submarine team behavioral toolkit rating scales

20.2.3 Electroencephalography

X-10 wireless headsets from Advanced Brain Monitoring Inc. were used for data collection. This wireless EEG headset system included sensor site locations F3, F4, C3, C4, P3, P4, Fz, Cz, and POz in a monopolar configuration referenced to linked mastoids; bipolar derivations were included, which have been reported to reflect sensorimotor activity (FzC3), workload (F3Cz, C3C4), and alpha wave components of the human mirror neuron system (Wang, Hong, Gao, & Gao, 2007). Embedded within the EEG data stream from each team member were eye blinks, which were automatically detected and decontaminated using interpolation algorithms contained in the EEG acquisition software (Levendowski et al., 2001). These eye-blink interpolations represented $\sim 5\%$ of the simulation time and in previous studies have not significantly influenced the detection of team neurophysiologic activities that occurred throughout the performances (Stevens et al., 2012). The EEG power values were computed each second at each sensor for the 1–40 Hz frequency bins by B-Alert Lab software.

20.2.4 Modeling Neurodynamic Symbol Streams

To generate neurodynamic symbols (NS) for the six-person navigation teams, each second, the power levels of one (of the 40) 1 Hz EEG frequency bin of a team member were equated with his or her own average levels over the task. This identified whether, at a particular time point, an individual team member was experiencing average (coded as 1), above average (coded as 3), or below average (coded as −1) levels of EEG power. The values for each person were combined each second into a vector, which was displayed as a histogram. For instance, the symbol in Fig. 20.2a represents a second when crew members 1 and 3 had below average EEG levels and the remaining crew had above average levels.

Generating the set of symbols over the entire performance (i.e., including briefing, scenario, and debriefing segments) provided neurodynamic models encompassing a comprehensive set of task situations/loads. Figure 20.2b shows the complete neurodynamic symbolic state space when each second of the performance was symbolically processed. Each NS situated the EEG power levels of each team member in the context of the levels of the other team members, and when the second-by-second symbols were aligned, the data stream contained a history of the team's neurodynamics. This history can be visualized by plotting the 25 NS each second where they can be related to training segments and activities (Fig. 20.2c).

A quantitative readout of this history could then be generated by calculating the Shannon entropy of the symbol distribution over a 100 s moving window

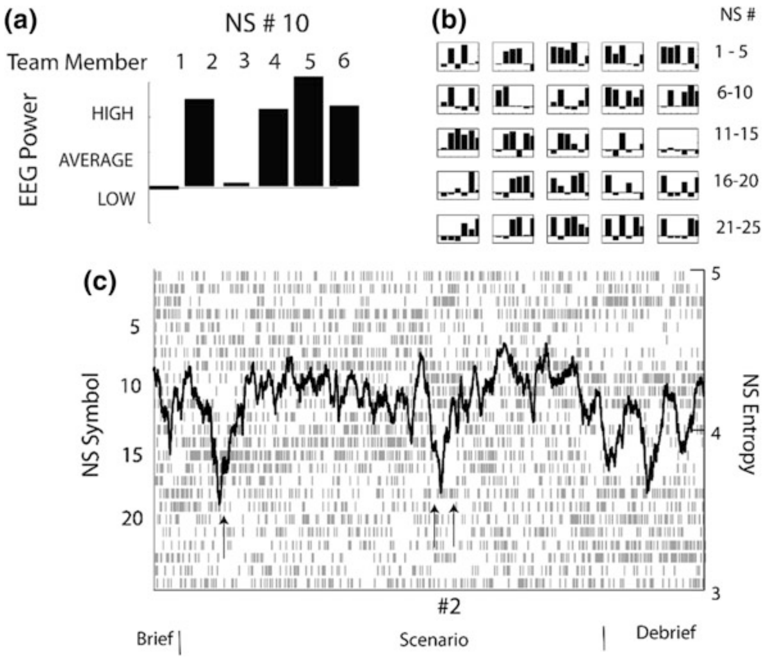


Fig. 20.2 Steps for extracting low-dimensional, single-trial neurodynamic organization information from the 10 Hz electroencephalography (EEG) levels of submarine navigation teams. **a** This symbol represents times when crew members 1 and 3 had below average 10 Hz EEG levels and the remaining crew had above average levels. **b** The 25-symbol state space is shown with the symbols assigned numbers in rows. **c** Each row represents the sequential expression of the 25 neurodynamic symbols (NS) from the 10 Hz frequency bin. These patterns are overlaid with a trace of the Shannon entropy of the NS symbol stream. The *single arrow* indicates when the crew had difficulty establishing the ship's position, and the *double arrows* indicate when the simulation was paused

(Shannon, 1951). Performance segments with restricted symbol expression like those indicated by the arrows in Fig. 20.2c had lower entropy levels, which is thought to reflect rigidity, whereas segments with greater symbol diversity had higher entropy, which is thought to reflect neurodynamic flexibility.

The goal of symbolic modeling was to measure the changing neurodynamic organizations of teams over different training segments and during realistic teamwork. As uniform models and scales are used for all teams and task segments, comparisons can be made across teams, across tasks, and over time (Fishel, Muth, & Hoover, 2007). The symbolic representations make the quantitation of the neurodynamic organizations explicit, and though the numeric aspects of each team member are less emphasized, their relationships are present in the symbol lookup table in Fig. 20.2b.

20.3 Results

The NS entropy levels for 12 SPAN teams were determined for the briefing, scenario, and debriefing segments across the 1–40 Hz EEG frequency bands (Fig. 20.3). The highest average NS entropy (i.e., the least team neurodynamic organization) occurred in the scenario segments, whereas significantly lower entropy levels (i.e., more team neurodynamic organization) were observed in the brief and debrief segments, $F(2, 11) = 3.52$, $p = 0.04$.

The NS entropy profiles were the highest at the lower (3–7 Hz) frequencies and progressively decreased toward the 40 Hz band. In each of the three training segments, there was also a pronounced decrease in NS entropy in the α region. When the symbol sequences were disrupted by randomization, the distinctiveness of the entropy profiles were lost.

The NS entropy–frequency profiles similar to that in Fig. 20.2c for the 10 Hz frequency bin were expanded to all 40 1 Hz frequency bins, and these neurodynamic organization maps showed the entropy (*z*-axis) at each frequency (*x*-axis) for each second (*y*-axis) of the performance.

The neurodynamic organization map for one of the least resilient teams is shown in Fig. 20.4; viewed from above, the darkened contours on the map showed the periods of decreased NS entropy, which were then aligned with different simulation events/team activities. This team was rated as having a low level of resilience, that is, unstressed battle rhythm (rating 1.0), where both evaluators indicated that leader presence was largely absent in this team. Commands were often informal and conversationally phrased or posed as a question. Task awareness was listed as being absent. This performance had the usual briefing and debriefing segments that

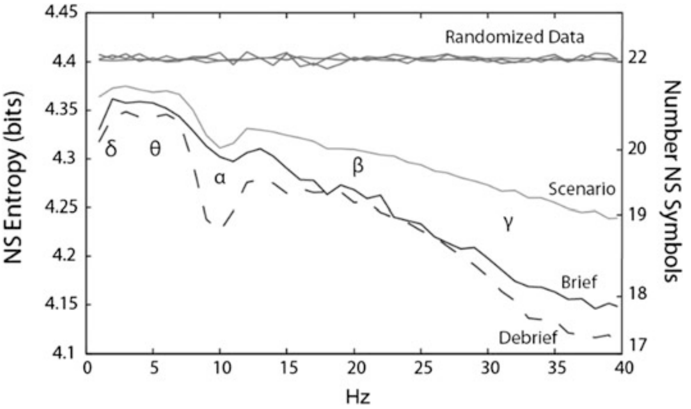


Fig. 20.3 Electroencephalography frequency profiles of team neurodynamic entropy. The neurodynamic symbol entropy streams from 12 submarine piloting and navigation performances were separated into the brief, scenario, and debrief segments, and the frequency–entropy profiles were generated. The *lines* labeled "Randomized Data" are the entropy profiles that resulted when the brief, scenario, and debrief symbol streams were randomized before calculating entropy

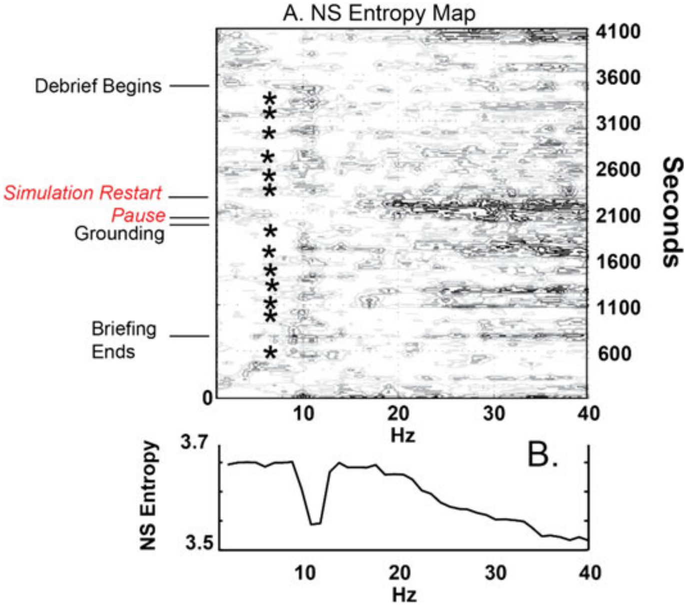


Fig. 20.4 **a** Time × Frequency × Entropy map was created for the 1–40 Hz frequencies of a submarine piloting and navigation team. Significant events are labeled to the *left*, and the *asterisks* (*) indicate the mark rounds calls. **b** The entropy values were averaged column-wise, creating a frequency–entropy histogram

bookended the scenario but was unusual in that midway through the performance, at ~1940s, the simulated submarine approached shoal water and grounded, a catastrophic event. The simulation was paused, the team was briefed by the instructor, and, at 2270 s, the submarine was repositioned in the simulation and the exercise continued.

Visual inspection of the NS entropy contours suggested close matches between the calls to mark rounds every 3 min and periods of minimum entropy in the 10 Hz frequency profile, suggesting that the team became more neurodynamically organized during this repeating activity. Across the 12 rounds cycles in the scenario, the NS entropy in the 10 s before the mark rounds call was significantly lower than that 70 s earlier, as the team began preparing to perform rounds or later (mark rounds = 3.45 ± 0.1 bits, 1 min to next round = 3.59 ± 0.1 bits, remaining seconds = 3.50 ± 0.1 bits), $H = 56.1$, df = 2, $p < 0.001$ (Kruskall–Wallis H test). As a control, similar comparisons at the 20 Hz frequency were not different, $H = 2.8$, df = 2, $p = 0.24$.

More variable periods of decreased NS entropy also occurred in the 20–40 Hz frequency bands (beta and gamma regions), the largest of which coincided with the simulation pause immediately after the grounding.

The previous data indicate that inter- and intrasegment neurodynamic synchronizations/organizations were frequent during SPAN teamwork, raising the question of whether these synchronizations also had significance in the broader context of team performance. Seven of the 12 SPAN team performances in this study met the criteria of having (a) been rated by two independent STBT evaluators, (b) the complete performances (i.e., the briefing, scenario, and debriefing segments) available for EEG modeling, (c) high-quality EEG (i.e., <15% eye-blink/muscle contamination) collected from at least five crew members, and (d) each of the training segments at least 500 s long.

The correlation between STBT observer ratings and the neurodynamic entropy of the entire performance (i.e., the brief, scenario, and debrief segments) was not significant, $r = -0.28$, $p = 0.53$. Correlations were repeated after separating the performance into the briefing, scenario, and debriefing segments. Between-group analysis of variance comparisons were significantly different, $F(2, 6) = 17.4$, $p < 0.001$, and a multiple comparisons analysis by Least Significances Differences indicated that the brief, scenario, and debrief segments differed at the 0.05 level.

During the briefing, there was a negative correlation, $r = -0.81$, $p < 0.005$, between the NS entropy and the STBT ratings, indicating that the more resilient teams were neurodynamically more organized than the less resilient teams. During the scenario, there was a positive correlation between STBT ratings and NS entropy, $r = 0.43$, $p = 0.04$, indicating that highly resilient teams were neurodynamically less organized than the less resilient teams. During the debriefing, the correlation was again negative, $r = -\ 0.36$, $p = 0.03$. The negative correlation means that higher STBT rating scores were correlated with lower NS entropy levels, that is, more synchronized and organized teams.

To situate the correlations into the EEG frequency spectrum, correlation and significance profiles were then constructed for each of the 40 EEG 1 Hz frequency bins for the briefing, scenario, and debriefing segments using the CzP0 EEG channel. As shown in Fig. 20.5, the correlations between entropy and STBT ratings were negative and significant at the $p < 0.01$ level in the ~20–40 Hz bins (β and γ

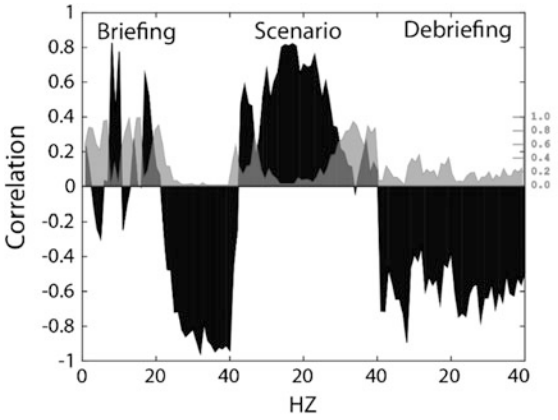


Fig. 20.5 Correlations between the Submarine Team Behavioral Toolkit evaluation scores and the electroencephalography power spectral density levels for the 1 Hz frequency bins from CzP0 ($n = 7$). The p-values for each correlation are shown in light gray. The briefing, scenario, and debriefing segments are labeled for each of the 40 1 Hz bins

rhythms) of the briefing segment. During the scenario segment, the sign of the NS entropy/STBT rating correlations reversed, with the most significant correlations, $p < 0.05$, between $\sim$10 and 20 Hz (β rhythms). Negative correlations were seen in the debriefing, although most were not significant at the 0.05 level.

20.4 Discussion

In this study, the linkages between the behavioral observations of evaluators and neurodynamic measures of teams performing submarine navigation tasks were explored. The approach taken was to identify extended periods (minutes) where the team members developed persistent neurodynamic organizations with regard to EEG power levels.

Most teams had characteristic NS entropy features, the first being the periods of lower NS entropy during the briefing and debriefing segments. This was not surprising as the teams are behaviorally the most organized during the debriefing, when all team members actively participate in the performance critique. The briefing segment is more a hybrid of the scenario and debriefing segments with periods of common discussion intermixed with individual instrument calibrations and small-group activities.

The neurodynamic synchronizations and organizations of teams were observed in most EEG frequency bands, with the possible exceptions of theta (θ) and delta (δ) regions. The neurodynamic organizations in the alpha (α) region dominated the NS entropy spectral profile for SPAN teams. The alpha band oscillations have known heterogeneity with regard to social coordination markers. The μ medial, the phi complex, and occipital α rhythms exist in the small frequency range of 9.5–13 Hz, with their amplitudes depending on whether the social coordination is intentional or incidental and whether the tasks are synchronic or diachronic.

Both varieties of these interactions would be expected in the SPAN task. Synchronic interactions dominate during the scenario, where information flows multidirectionally across all members of the crew, whereas during the debriefing segment, only one person generally speaks at a given time (i.e., diachronic interaction). These may in part account for the entropy differences between these segments seen in Fig. 20.3. The scenario–debriefing differences in NS entropy in the alpha region might also result from increased/prolonged periods of alpha suppression resulting from the increased task requirements in the scenario (Klimesch, Sauseng, & Hanslmayr, 2007).

While the alpha NS entropy dominated the NS EEG spectral profile, the correlations between EEG frequency and STBT ratings in different segments (Fig. 20.5) suggest they may be less important for distinguishing between high- and low-resilience teams as alpha NS entropy levels were poorly correlated with team resilience. This may in part be due to the central role of alpha NS organizations during the taking of rounds, which is a periodic and routine activity that all crew members had extensive experience with. The subjects studied were candidates in advanced training and already had several years' operational practice performing the rounds routines, and at least one of the social coordination markers in the alpha region (right mu) decreases when people memorize routine behaviors of others (Tognoli & Kelso, 2013).

The NS entropy decreases in the γ region are more enigmatic as social coordination markers have not yet been described in this region. In individuals however, α, β, and γ oscillations interact during working memory manipulations (Roux & Uhlhaas, 2014). In this regard it is interesting that periods of γ synchronization were often observed in association with oscillations in α and β bands as well.

The periods of increased team neurodynamic organization during the scenarios were concurrent with "periods of interest" for the team. The clearest example was the simulation pause in Fig. 20.4, where the NS entropy levels dropped to those seen in debriefings.

Generally, the higher performing teams had fewer periods of reorganization and/or periods of smaller duration or magnitude during the scenario (Stevens, Gorman, Amazeen, Likens, & Galloway, 2013). This observation was confirmed by correlation analysis between team synchrony and STBT ratings. What was unexpected from these analyses was the negative correlation between team synchronization and evaluator ratings in the briefing segment. This relationship suggests that the more cognitively organized a team was during the briefing, the better they would perform on the task. If larger scale fluctuations indeed relate to the need for increased team organization, then by identifying significant periods of team reorganization, instructors could advantageously target discussions and future training activities to develop team skills in these areas and to objectively follow team improvement over time. Neurodynamic measures may also have utility for determining when a team is becoming brittle or "drifting into danger." Detection of team breakdowns can be difficult because of the subtle onset and multiplicity of causes before a critical transition toward failure occurs (Woods & Hollnagel, 2006). Although team breakdown can be perceived as a sudden event with a dramatic loss of effectiveness, more often, this decrease in performance is a gradual or incremental process (Rankin, Lunderg, Woltjer, Rollenhagen, & Hollnagel, 2014). Better determining when the team began reorganizing would be a step forward toward understanding the antecedent events to difficulties and toward developing strategies to mitigate against them in the future, perhaps in real time.

The symbolic representations and modeling, though useful as intermediate representations between microneural events and macro-observational ratings, are not without limitations, because it is uncertain what exactly is being measured cognitively. To some extent, this is not surprising, as details of teaming are poorly understood in the tens of seconds to extended minutes timescale.

Recently, similar findings have been seen with health care teams when correlations were performed between neurodynamic entropy levels and TeamSTEPPS ratings (Stevens, Galloway, Gorman, Willemsen-Dunlap, & Halpin 2016). The negative correlation in the briefing and positive correlation in the scenario, along with similar frequency characteristics of both the submarine and health care team correlations, suggest that the underlying construct might be common to this class of team activity. A better understanding of these meanings can be approached by more detailed modeling across multiple sensor locations or spatially independent components (Onton, Westerfield, Townsend, & Makeig, 2006).

Finally, collapsing the team measures into a single data stream simplifies linking them with other data streams of team performance (speech, gestures, etc.). In this way, Gorman et al. (2015) have shown novice–expert differences in the correlational time lags associated with team neurodynamics and team speech.

Acknowledgements This work was supported in part by The Defense Advanced Research Projects Agency under contract number(s) W31P4Q12C0166, and NSF SBIR grants IIP 0822020 and IIP 1215327. The views, opinions, and/or findings contained are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. This work was supported by work unit number F1214. JL is an employee of the U.S. Government. This work was prepared as part of my official duties. Title 17 U.S.C. 105 provides that 'copyright protection under this title is not available for any work of the United States Government.' Title 17 U.S.C. 101 defines U.S. Government work as work prepared by a military service member or employee of the U.S. Government as part of that person's official duties. The study protocol was approved by the Naval Submarine Medical Research Laboratory Institutional Review Board in compliance with all applicable Federal regulations governing the protection of human subjects.

References

Anders, S., Heinzle, J., Weiskopf, N., Ethofer, T., & Haynes, J. (2011). Flow of affective information between communicating brains. *Neuroimage, 54,* 439–446.

Baker, D. P., Amodeo, A. M., Krokos, K. J., Slonim, A., & Herrera, H. (2009). Assessing teamwork attitudes in healthcare: Development of the TeamSTEPPS® teamwork attitudes questionnaire. *Quality and Safety in Health Care, 19*, e49. doi:10.1136/qshc.2009.036129

Battaglia, F. P., Sutherland, G. R., & McNaughton, B. L. (2004). Local sensory cues and place cell directionality: Additional evidence of prospective coding in the hippocampus. *Journal of Neuroscience, 24,* 4541–4550.

Bonnefond, M. & Jensen, O. (2015). Gamma activity coupled to alpha phase as a mechanism for top-down controlled gating. *PLOS One, 10*(6), e0128667. doi:10.1371/journal.pone.0128667

Buzaki, G. (2006). *Rhythms of the brain*. New York: Oxford University Press.

Caetano, G., Jousmaki, V., & Hari, R. (2007). Actor's and observer's primary motor cortices stabilize similarly after seen or heard motor actions. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 9058–9062.

Cooke, N. J., Gorman, J. C., & Kiekel, P. (2008). Communication as team-level cognitive processing. In M. P. Letsky, N. W. Warner, S. M. Fiore, & C. A. P. Smith (Eds.), *Macrocognition in teams* (pp. 51–64). Burlington, VT: Ashgate.

Dmochowski, J. P., Sajda, P., Dias, J., & Parra, L. (2012). Correlated components of ongoing EEG point to emotionally laden attention—A possible marker of engagement? *Frontiers in Human Neuroscience, 6*, Article 112.

Fishel, S. R., Muth, E. R., & Hoover, A. W. (2007). Establishing appropriate physiological baseline procedures for real-time physiological measurement. *Journal of Cognitive Engineering and Decision Making, 1,* 286–308.

Flack, J. C. (2012). Multiple time-scales and the developmental dynamics of social systems. *Philosophical Transactions of the Royal Society B, 367*, 1802–1810. doi:10.1098/rstb.2011.0214

Gorman, J., Martin, M., Dunbar, T., Stevens, R. H., Galloway, T. L., Amazeen, P., et al. (2015). Cross-level effects between neurophysiology and communication during team training. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*, 181–199. doi:10.1177/0018720815602575

Harmony, T. (2013). The functional significance of delta oscillations in cognitive processing. *Frontiers in Integrative Neurosciences, 7*, Article 83. doi:10.3389/fnint.2013.00083

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Inter-subject synchronization of cortical activity during natural vision. *Science, 303,* 1634–1640.

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: The inhibition-timing hypothesis. *Brain Research Reviews, 53,* 63–88.

Levendowski, D. J., Berka, C., Olmstead, R. E., Konstantinovic, Z. R., Davis, G., Lumicao, M. N., et al. (2001). Electroencephalographic indices predict future vulnerability to fatigue induced by sleep deprivation. *Sleep, 24*(Abstract Suppl.), A243–A244.

Lisman, J. E., & Jensen, O. (2013). The theta-gamma code. *Neuron, 77,* 1002–1016.

Menoret, M., Varnet, L., Fargier, R., Cheylus, A., Curie, A., desPortes, V., et al. (2014). Neural correlates of non-verbal social interactions: A dual-EEG study. *Neurophyschologia, 55,* 85–91.

Nummenmaa, L., Gleran, E., Viinikainen, M., Jaaskelainen, P., Hari, R., & Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 9599–9604.

Oberman, L. M., Pineda, J. A., & Ramachandran, V. S. (2007). The human mirror neuron system: A link between action observation and social skills. *Social Cognitive and Affective Neuroscience, 2,* 62–66.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Research, 13,* 171–175.

Onton, J., Westerfield, M., Townsend, J., & Makeig, S. (2006). Imaging human EEG dynamics using independent component analysis. *Neuroscience and Behavioral Reviews, 30,* 808–820.

Palva, S., & Palva, J. M. (2007). New vistas for α-frequency band oscillations. *Trends in Neuroscience, 4,* 150–158.

Pineda, J. A. (2008). Sensorimotor cortex as a critical component of an "extended" mirror neuron system: Does it solve the development, correspondence, and control problems in mirroring? *Behavioral and Brain Functions, 4,* 47–63.

Rankin, A., Lunderg, J., Woltjer, R., Rollenhagen, C., & Hollnagel, E. (2014). Resilience in everyday operations: A framework for analyzing adaptations in high-risk work. *Journal of Cognitive Engineering and Decision Making, 8*(1), 78–97.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience, 9,* 661–670.

Roux, F., & Uhlhaas, P. (2014). Working memory and neural oscillations: Alpha-gamma versus theta-gamma codes for distinct WM information? *Trends in Cognitive Sciences, 18,* 16–25.

Schippers, M., Roebroeck, A., Renken, R., Nanetti, L., & Keysers, C. (2010). Mapping the information flows from one brain to another during gestural communication. *Proceedings of the National Academy of Sciences of the United States of America, 107,* 9388–9393.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal, 30,* 50–64.

Shockley, K., Santana, M.-V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 326–332.

Stephens, G., Silbert, L., & Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 14425–14430. doi:10.1073/pnas.1008662107

Stevens, R. H., & Galloway, T. (2014). Toward a quantitative description of the neurodynamic organizations of teams. *Social Neuroscience, 9,* 160–173.

Stevens, R. H. & Galloway, T. (2015). Modeling the neurodynamic organizations and interactions of teams. *Social Neuroscience, 11*, 123–139. doi:10.1080/17470919.2015.1056883

Stevens, R. H., Galloway, T., Wang, P., & Berka, C. (2011). Cognitive neurophysiologic synchronies: What can they contribute to the study of teamwork? *Human Factors, 54*, 489–502. doi:10.1177/0018720811427296

Stevens, R. H., Galloway, T., Wang, P., Berka, C., Tan, V., Wohlgemuth, T., et al. (2012). Modeling the neurodynamic complexity of submarine navigation teams. *Computational and Mathematical Organization Theory, 19,* 346–369.

Stevens, R. H., Gorman, J. C., Amazeen, P., Likens, A., & Galloway, T. (2013). The organizational dynamics of teams. *Nonlinear Dynamics, Psychology, and Life Sciences, 17*(1), 67–86.

Stevens, R. H., Galloway, T.L., Gorman, J., Willemsen-Dunlap, A., Halpin, D. (2016). Toward objective measures of team dynamics during healthcare simulation training. Presentation #89, *International Symposium on Human Factors and Ergonomics and Health Care*, San Diego, CA.

Tognoli, E. & Kelso, J. A. (2013). *The coordination dynamics of social neuromarkers.* Retrieved from arXiv database (Preprint No. 1310.7275).

Tognoli, E., Lagarde, J., De Guzman, G. C., & Kelso, J. A. S. (2007). The phi-complex as a neuromarker of human social coordination. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 8190–8195.

Wang, Y., Hong, B., Gao, X., & Gao, S. (2007). Design of electrode layout for motor imagery based brain-computer interface. *Electronics Letters, 43,* 557–558.

Woods, D. & Hollnagel, E. (2006). Resilience engineering concepts. In E. Hollnagel, D. D. Woods & N. Leveson (Eds.), *Resilience engineering: Concepts and precepts* (pp. 1–6). Burlington, VT: Ashgate.