Language Learning & Language Teaching 12

Inference and Generalizability in Applied Linguistics

Multiple perspectives

Edited by
Micheline Chalhoub-Deville
Carol A. Chapelle
Patricia Duff

John Benjamins Publishing Company

Inference and Generalizability in Applied Linguistics

Language Learning and Language Teaching

The *LL&LT* monograph series publishes monographs as well as edited volumes on applied and methodological issues in the field of language pedagogy. The focus of the series is on subjects such as classroom discourse and interaction; language diversity in educational settings; bilingual education; language testing and language assessment; teaching methods and teaching performance; learning trajectories in second language acquisition; and written language learning in educational settings.

Volume 12

Inference and Generalizability in Applied Linguistics: Multiple perspectives
Edited by Micheline Chalhoub-Deville, Carol A. Chapelle and Patricia Duff

Inference and Generalizability in Applied Linguistics

Multiple perspectives

Edited by

Micheline Chalhoub-Deville
University of North Carolina, Greensboro

Carol A. Chapelle
Iowa State University

Patricia Duff
University of British Columbia

John Benjamins Publishing Company
Amsterdam/Philadelphia

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Inference and Generalizability in Applied Linguistics : Multiple perspectives / edited by Micheline Chalhoub-Deville, Carol A. Chapelle and Patricia A. Duff.
p. cm. (Language Learning and Language Teaching, ISSN 1569–9471 ; v. 12)
Includes bibliographical references and indexes.
1. Applied linguistics--Research. 2. Inference. 3. Reliability. I. Chalhoub-Deville, Micheline. II. Chapelle, Carol. III. Duff, Patricia A. IV. Series.

P129.I455 2006
418.072--dc22 2005055573
ISBN 90 272 1963 X (Hb; alk. paper)
ISBN 90 272 1964 8 (Pb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

II. Discussion

Drawing the line

The generalizability and limitations of research in applied linguistics

Micheline Chalhoub-Deville
University of North Carolina, Greensboro

Research in applied linguistics investigates a range of complex issues and consequently it relies on equally complex approaches to research methodology. Methods for obtaining insights and evidence about language issues critically affect the outcomes of research and are therefore an important object for discussion and reflection. For a number of years, the joint sessions of the International Language Testing Association (ILTA) and the American Association for Applied Linguistics (AAAL) have provided a forum for such discussion and reflection. Many of the chapters in this volume originated in the session at the 2002 meeting of the AAAL in Salt Lake City. Following tradition for the joint ILTA-AAAL session, this one brought together researchers working in different areas of applied linguistics to share their views and debate issues.

The topic for the 2002 meeting, "Drawing the line: The generalizability and limitations of research in applied linguistics," asked participants to explicate the perspectives underlying their approaches to data collection, analysis, interpretation of results, and generalizability of findings. In particular, participants were to discuss their views on the meaning and relevance of (1) dependability in data collection and analysis, (2) inferences made on the basis of observations, and (3) generalization of research results. Participants, most of whom have contributed to this volume, were asked to address the following questions with respect to their specific area of investigation:

1. How does research in your area address the dependability and generalizability of elicited judgments and observations?

2. How does research in your area deal with the appropriateness of inferences made based on observed learner performances?
3. How are issues of dependability, generalizability, and appropriateness of inferences dealt with in diverse research paradigms prevalent in your area?

The three terms dependability, generalizability, and inference are obviously central to these questions. Although the purpose of this volume is to define and discuss these terms from a variety of perspectives, I offer a brief orientation to each of the concepts as a point of departure. My own education, background, and biases in language assessment are readily evident in both the concepts selected for discussion and how I define them.

Dependability

Dependability or reliability, in a broad sense, refers to the consistencies of data, scores, or observations obtained using elicitation instruments, which can include a range of tools from standardized tests administered in educational settings to tasks completed by participants in a research study. From an assessment perspective, the random variation obtained from performance is called "error," which limits the degree of reliability, or dependability, of the results obtained from the instrument. One of the preoccupations of researchers in assessment is identify sources of error variance that might impinge on assessment results.

One such source of error is the elicitation method itself. As a researcher, how do I know that the method or instrument I am using to collect my data will produce dependable results? The fact that I, as a researcher, thought the instrument would render results deemed trustworthy does not necessarily mean that it will. Evidence should be provided to document the dependability of results obtained from any instrument used in our investigations.

A second potential source of error is the number of elicitations. There is a potential danger in getting information from only one observation. In principle, the more observations generated, the more confidence one can have in the dependability of results.

A third potential source of error can be attributed to the influence of the interlocutor or observer involved in the elicitation. The rater/researcher judging the quality of a performance plays a significant role in the results obtained and therefore can introduce error. In summary, researchers need to identify and document all factors that play a salient role in data collection. The point of identifying these sources of variability is to minimize the error they pro-

duce and estimate their magnitude so that results can be interpreted and used appropriately.

Generalizability

Generalizability refers to the extent to which research results can justifiably be applied to a situation beyond the research setting. What are the characteristics of the domain, e.g., the learners and tasks, to which the results apply? The characteristics of the learners under investigation and the larger group to which we want to generalize or transfer our results are issues to consider when discussing generalizability. The comparability of learners in a study to those who were not included should be discussed in terms of the applicability of research results. The methods used for eliciting samples from language learners play a critical role in terms of generalizability because elicitation methods affect the type of data obtained. We, as researchers, need to be aware of instrument effect on the phenomenon under examination and therefore on the resulting observations; findings based on one method/task cannot be assumed to generalize to performance on similar/dissimilar tasks.

Inferences

In language assessment, inference refers to the connection that the researcher makes between observations of learners' performance and interpretations about the meaning of that performance. The process of inference is not automatic or independently objective – it requires justification backed by evidence. Also critical to the construction of inferences is who is making these inferences and for what purpose. For example, we can argue that test developers and researchers are not the only ones who provide interpretations of scores; test takers and users also interpret meaning from scores. The link between data generated and the interpretations and intended uses requires making a case based on the convergence of theoretical arguments and empirical evidence from multiple sources.

The three concepts – dependability, generalizability, and inference – are dealt with implicitly or explicitly in any research undertaken in applied linguistics. As already indicated, however, one could make the case that the questions, as defined and posed here, are indicative of a particular research paradigm or/and perspective. One may argue that the concepts focused upon in these

questions may not be considered particularly relevant from certain research perspectives, or they may be emphasized in different ways. Another consideration is the concern that these questions and concepts stand in opposition to each other. As such, what is the appropriate balance of these concepts? Where are we to draw the line on how much attention to pay to each of these concepts? This volume begins to explore how these concepts come into play differently across various research perspectives. Such cross-paradigm communication is ultimately in the best interest of applied linguistics as a discipline and will hopefully engender increased confidence in the claims made on the basis of our field's research.

Chapters in this volume

The chapters in this volume are organized into two sections. The first section consists of chapters that address the three questions listed above – each chapter received comments from the editors and was revised. The second section contains discussion chapters. Discussants read and offered commentary on the chapters presented in the first section – discussants had a free hand to comment on and discuss these chapters.

The first chapter, "Old and new thoughts on test score variability: Implications for reliability and validity" by Craig Deville & Micheline Chalhoub-Deville, focuses on the meaning of variability when interpreting scores from an assessment. The second chapter, "Validity and values: Inferences and generalizability in language testing" by Tim McNamara provides extensive discussion of the three terms and issues based on work in language assessment. In the third chapter, Carol Chapelle applies the three concepts to the problem of theory development and assessment of the L2 lexicon in a chapter entitled "L2 vocabulary acquisition theory: The role of inference, dependability and generalizability in assessment."

The fourth chapter, "Beyond generalizability: Contextualization, complexity, and credibility in applied linguistics research" by Patricia Duff provides a transition by explaining the concerns of quantitative researchers and contrasting these with the values and priorities of qualitative researchers. Next, Merrill Swain examines the qualitative methodology of verbal protocols as a means for eliciting SLA data in terms of data interpretation and generalization in her chapter, "Verbal protocols: What does it mean for research to use speaking as a data collection tool?" Diane Larsen-Freeman's chapter, "Functional grammar: On the value and limitations of dependability, inference, and generalizability"

explores dependability, generalization, and inference in the study of functional grammar. Finally, Numa Markee looks at how these terms apply in conversation analysis research with his chapter, "A conversation analytic perspective on the role of quantification and generalizability in second language acquisition."

The three discussion chapters in the second section identify cross-paradigmatic issues and extend the discussion to larger issues in applied linguistics research. Lyle Bachman's chapter, "Generalizability: A journey into the nature of empirical research in applied linguistics," draws on perspectives from the philosophy of science to explain differing views across the chapters. Susan Gass explores the importance of generalizability for interpretations made about learning in her chapter, "Generalizability: What are we generalizing anyway?" Finally, Nancy Hornberger discusses the points of disagreement and synergy across the chapters in her contribution, "Negotiating methodological rich points in applied linguistics research: An ethnographer's view."

By moving the ideas raised at the AAAL conference venue into print, the editors hope that the "rich points," as Hornberger put it, will appear more salient to researchers in applied linguistics and that these points will serve as an impetus for continued discussion and development in our profession.

PART I

Perspectives on inference and generalizability in applied linguistics

Old and new thoughts on test score variability

Implications for reliability and validity

Craig Deville and Micheline Chalhoub-Deville
University of North Carolina, Greensboro

The paper discusses the variability of test takers' performances across different language tasks. We concentrate attention on this aspect of variability because inconsistent achievement by test takers across tasks has emerged as a significant threat to reliability and validity. The first section of the chapter addresses how language testers have traditionally conceptualized and measured variability, while the second part advocates an alternate way of thinking about the issue. Our intent should not be construed as a defense of or apology for the dominant paradigm employed by many language testers. Instead, we hope to problematize our standard thinking of concepts such as variability and error, concepts dear to the heart of conventional testers but sometimes anathema to those of a different epistemological bent.

Introduction

Language testers are typically interested in meaningfully measuring the second language (L2) proficiency of test takers and making inferences from the test scores and/or observed behaviors to a test taker's ability to use language in a particular context or for an identified purpose. We language testers wish to make generalizations – sometimes across test takers, other times across items or tasks, and/or at still other times across occasions or contexts. Making inferences and generalizations involves us in issues of reliability and validity. Such a perspective and approach to testing is rather conventional, based on well-known psychometric principles. The approach has been labeled positivist (McNamara 2001). One can dispute the positivist label (Yu 2003), but McNamara's argu-

ment as to language testers' over-reliance on a particular paradigm of knowledge, research, and values is right on target.

The present authors carry a load of 'positivist' baggage and will therefore address issues of reliability and validity from that standpoint in the first part of this chapter. Our emphasis will be, from a measurement perspective, on the inferences and dependability/generalizability of L2 test scores across test items/tasks and, tangentially, extrapolations to real-life situations in which one uses their L2. Like others (e.g., Bachman & Palmer 1996; Chapelle 1999) we incorporate issues of reliability under the umbrella of test score validation and consider reliability evidence as a necessary component of a validity argument. The reader is thus encouraged to keep in mind that validating the interpretation and use of L2 test scores underlies any discussion of what constitutes reliability.

We will use the three terms *reliability*, *dependability*, and *generalizability* as synonymous unless specifically indicated otherwise. We use these terms to refer to the consistency of test scores across various facets of measurement such as items, tasks, raters, occasions, etc. We use the terms *infer* and *inference* to indicate extrapolation from a test taker's performance on a test to her/his performance in a real-life situation. We will *not* discuss the representativeness or generalization of samples of test takers to larger populations. Although this is an important consideration when evaluating certain inferences about test scores, we see it as an issue related to sampling procedures and not directly to score reliability.

In this chapter our focus will be primarily on the variability of test takers' performances across different language tasks. We choose to concentrate attention on this aspect of variability because inconsistent achievement by test takers across tasks has emerged as the most significant threat to the reliability and validity of test scores from performance assessments. The first section of the chapter addresses how language testers have traditionally conceptualized and measured variability. The second part advocates an alternate way of thinking about several of these issues, predominately how we think about test score variability.

The concept of variability is critical to any discussion of reliability, and by extension, of validity. Our intent in this chapter should not be construed as a defense of or apology for the dominant paradigm employed by many language testers. Instead, we hope to problematize our standard thinking of concepts such as variability and error, concepts dear to the heart of conventional testers but sometimes anathema to those of a different epistemological bent.

In conclusion, the present chapter discusses common practices and considerations in the language testing field, including some concerns and limitations of standard practices, and suggests alternative research approaches that may enrich our understanding of the L2 construct. (The term, L2 construct, is used as a placeholder throughout the chapter to represent any language construct of interest to researchers.) The authors draw on literature outside of the applied linguistics domain, e.g., from the areas of general measurement and developmental psychology, in order to inject some different and fresh perspectives on the L2 construct investigation.

The roles of language testers

Language testers usually wear two hats. First and foremost they are testers. One might even say, applied testers. They are often commissioned to develop a L2 test because an organization or institution has a need to distinguish among examinees or candidates with respect to their L2 proficiency. The commissioner of the test is often the one(s) who largely determines the purpose and context of the assessment, the decisions to be made based on scores, and the numerous administrative and operational constraints on test development and research. Within this framework language testers have investigated numerous research questions and advanced the knowledge of the field with respect to the L2 construct. For the sake of the present discussion, however, the authors classify language testers as both applied and basic researchers and examine the respective orientations embodied in these two roles.

Sometimes, indeed rarely, a language tester might have the luxury of wearing her research hat without having to worry about the many stones in her pockets, pulling her under. To belabor the metaphor, the numerous stones or applied/pragmatic constraints on the tester prevent her from undertaking unencumbered construct research because the heavy realities just barely allow her to keep her head above water. Nevertheless, we language testers can and should think beyond the impediments of conventional testing and consider what we as researchers would like to investigate. But first, beginning with our applied tester's hat, we will focus on some important notions that influence how we typically think about reliability, while threading considerations of validity throughout our discussion.

Language testers as testers

Measurement professionals – language testers included – are expected to adhere to agreed upon standards that inform sound and ethical test practices. These are commonly known as the *Standards* (see AERA, APA, & NCME 1999). The *Standards*, reflecting the fact that psychometrics, assessment practices, legal decisions, and social values are constantly evolving, have been revised and re-published approximately every decade since their debut in the 1950s. Each edition, however, has emphasized the importance of reliability as a necessary component when evaluating tests and testing practices. The most recent *Standards* (1999) state that:

> The usefulness of behavioral measurements presupposes that individuals and groups exhibit some degree of stability in their behavior. However, successive samples of behavior from the same person are rarely identical in all pertinent respects. An individual's performances, products, and responses to sets of test questions vary in their quality or character from one occasion to another, even under strictly controlled conditions. This variation is reflected in the examinee's scores. (p. 25)

The assumption of a stable construct within individuals is a cornerstone when discussing traditional notions of reliability and validity. As testers, whose instruments are used to make decisions about test takers, we need to be concerned with how stable and/or how variable our test scores are.

Variability and error

The concepts of variability and error are crucial to any conventional discussion of reliability, so it is important to understand what we mean by these terms. Traditionally, second language acquisition (SLA) researchers (e.g., Ellis 1994; Tarone 1988) have examined variability of performance *within* an individual confronted with various language tasks in order to study task variables. On the other hand, language testers have examined variability of test scores *across* examinees so as to arrive at decisions with respect to the person's language ability, as inferred from the test performance. Many of us who are involved with language testing and measurement have a somewhat ambivalent view of the variability we find in test takers' scores. Assuming our interest is concerned with individual differences, we testers like to think that 'good' (i.e., relevant or useful) variability is what differentiates test takers from each other with respect to their language ability, or construct of interest. 'Bad' (or irrelevant, confounding) variability, on the other hand, is considered error, and it can discredit our faith in the test scores as trustworthy indicators of the differences among test

takers. This thinking, in a nut shell, represents the conventional, psychometric approach to conceptualizing and analyzing the reliability of test scores. The more relevant variability we have relative to the irrelevant, the higher – and better – the reliability estimate. That is, we say a set of scores are reliable when we can ascertain that differences among examinees are due to individual differences in the L2 construct, e.g., ability to read academic texts, and not due to measurement error, e.g., test environment conditions at a particular setting.

One of the primary intended outcomes of administering a test to a group of examinees is our ability to separate the test takers based on their scores. We might separate them along some continuum or score scale or we may separate them into categories, e.g., proficiency levels or pass-fail status. The point is that we use the test scores, in some fashion, to order our test takers so that we can infer which individuals are more proficient in the language, or in certain aspects of the language, and which are less so.

Many types of decisions are made based on test scores, i.e., on the inferences we make regarding the test taker's language ability. We assign grades/marks, we decide which students gain admission to our universities, we place students into certain courses, we select certain candidates for jobs, etc. For such decisions affecting test takers to be considered fair, we must, among other things, be able to depend on the scores as accurate indicators of the test takers' ability, especially when these decisions involve an inference of one examinee's ability relative to another's. A reliability estimate tells us something about how consistent or dependable our test scores are, and by extension, how much faith we can have in – how valid are our – subsequent inferences and decisions stemming from those scores.

As straightforward as the concept of reliability is, it becomes complicated when we attempt to define and explicate what comprises good and bad variability (and devise methods to quantify these). Much of the complication occurs because we have not adequately specified what constitutes good and bad variability. Too many test developers and researchers are willing to leave reliability to the psychometrician – or adopt whatever is made available by the psychometrician – who might simply calculates the prototypical reliability estimate, coefficient alpha. One issue here is that alpha may not be the appropriate estimate of reliability. But more importantly, the test developer/researcher and the psychometrician should consider how the assessment procedures might influence and/or determine both differences in the scores related to the construct and error in the test scores. For example, unless the different sources of test score variability are properly identified, as discussed in the following section,

it is impossible to accurately quantify how much variance can be attributed to the L2 construct of interest.

Conceptualizing reliability

Psychometricians have devoted extensive work over the decades to derive appropriate models for estimating reliability (see Feldt & Brennan 1989), and virtually all of these models are based on differing conceptualizations of what constitutes variability in test takers' performances that can be attributed to the construct of interest, and what constitutes variability due to error. It is not our intent here to present different reliability models, but instead to draw attention to situations where language test developers should consider what might lead to variability in test takers' scores.

As mentioned above, the most widely used indicator of reliability, i.e., alpha, may not always be the most appropriate estimate (Hogan, Benjamin, & Brezinski 2000). Depending on a number of factors with respect to how a given test is constructed (discussed below under the rubric of testlets), alpha may under- or over-estimate reliability. We language test developers should be aware of these factors where applicable and discuss them with the psychometrician so that reliability can be investigated and established appropriately. To illustrate this point, we consider reliability issues associated with testlets, which are quite common in language testing.

A testlet is a bundle of items whereby the items are linked because they belong to the same reading or listening passage, because they have the same format, among other reasons (Lee, Brennan, & Frisbie 2000). There are numerous approaches to estimating reliability when testlets are involved (e.g., Feldt 2002; Lee, Dunbar, & Frisbie 2001), but the point is that alpha is probably not an appropriate estimate because there is likely some dependency among items within bundles. Furthermore, of concern is not simply the use of a different estimation technique, but the consideration of how a testlet is defined with respect to the test content. For example, reading items may belong to a particular passage, but they might also be 'bundled' according to subskill (e.g., identifying author's purpose), type of passage (e.g., fiction or narrative), or other content strata (Lee 2002). And, however we parcel our items, do we consider them to represent a random sample of possible, say reading passages – a very plausible approach – or do they represent a fixed selection, e.g., reading passages covering science topics only? These substantive considerations are likely spelled out in the test specifications and can provide important guidance for estimating reliability.

Very much related to the use of testlets is the use of multiple item formats within a test, especially when these vary in length and their contribution to a total score (Feldt & Charter 2003; Qualls 1995). It is not uncommon for a language assessment, especially classroom assessments and placement exams, to contain some selected response items (e.g., multiple-choice), short answer questions, and even an extended question. Moreover, widespread practice is to award a different number of points for the various test parts. For example, a university placement examination may consist of 10 multiple-choice reading questions and a writing task worth 20 points. With such tests, some consideration should be given to how the different sections likely contribute to variability in 'true' scores and error. With the above mentioned placement test, it is clear that the two sections very likely contribute differentially to the variability of test scores. We might, then, further deliberate whether separate scores should be given for each test section, or what a composite score is intended to reflect. Again, these are all important deliberations so that we can accurately estimate reliability.

Another test format/method we language testers employ are so-called integrative tasks (see Oller 1979), such as cloze tests, C-tests, recall protocols, etc., which entail different challenges when estimating reliability (Deville & Chalhoub-Deville 1993). A careful investigation into how an 'item' is defined for scoring purposes (e.g., exact word cloze or recalled words versus propositions) and the interrelationship(s) of the items and/or item bundles is warranted. These issues have significant implications for how we measure reliability. Once again, the unquestioning adoption of traditional reliability estimates obviates the need to consider critical aspects of test construction, tied to test content, which should inform us when we estimate test score reliability.

The foregoing description of testing practices demonstrates the interrelatedness of test construction and content with how we measure reliability. These practices and considerations stem from the applied work we language testers do as testers. Moreover, they are based on the traditional operationalization of reliability found in both the general measurement and language testing literature. Moss (1992, 1994), however, offering an interpretive or hermeneutic paradigm, challenges the notion that traditional reliability is the *sine qua non* for good measurement and hence for any claim to validity as well. In a similar vein, Nesselroade and Ghisletta (2000), who work in the field of developmental psychology, argue that the fundamental perspective that constructs are stable entities – a premise they dispute, but one that is presupposed in the 1999 *Standards* (see citation above) – has influenced how we practice measurement, indeed how we conceptualize reliability and validity. Reliability estimates based

on test-retest correlations, a notion that forms the basis of classical test theory, obviously presumes a static construct within individuals. Finally, within the SLA and language testing fields, Swain (1993) has also argued that traditional notions and practices of psychometrics may not be appropriate or serve us well when we define the language construct and performance in terms of co-construction (see below).

In the next section we will address some alternative ways we might think about several of these standard testing concepts. We will look at language testers as basic researchers, freed from some of the pragmatic constraints imposed by the applied framework.

Language testers as researchers

What are the issues that language testers as researchers might investigate if they needn't be so concerned with generalizing test scores? In this section we continue to examine aspects of variability and error conceptualization as they pertain primarily to tasks. We note the consistent findings in the L2 and the general measurement fields with regard to an interaction between persons, i.e., test takers, and tasks and we consider the conventional interpretation of this variability as error/noise. We suggest that in our role as language testing researchers we can sidestep some of the usual thinking about variability and error and explore new ways of examining variability. We make a case that person x task variability is not to be relegated to nuisance or error variance but should be regarded as being at the heart of the study of the L2 construct. We then suggest how this conceptualization of interaction and variability can be examined within the context of a 'theory of context.' Being applied testers at heart, we maintain that this unconventional approach can inform both test score interpretation, i.e., what test takers' scores mean in regard to our construct definition, and test use, i.e., what decisions are made based on test takers' scores (Messick 1989).

Test method variability and the L2 construct

Many inferences about test takers are essentially generalizations/predictions about their performance in non-testing situations. We typically take a snapshot picture of a test taker's language ability based on the person's performance on a limited number and type of tasks. As practitioners, we would like to know that if we were to examine these same test takers again, we would obtain a consistent, replicable ranking of scores. We would like to be able to tell how reliable and/or generalizable the ranking of those scores is.

From both the language testing literature (Chalhoub-Deville 1995a, 1995b; Lee, Kantor, & Mollaun 2002; Shohamy 1984) and general measurement studies (Brennan 1992, 2001; Linn & Burton 1994; Shavelson, Baxter, & Goa 1993) we know that, even though our test items, tasks, or measurement methods purportedly represent the same content/construct, test takers perform differentially across these measures. This is typically referred to as a method or an interaction effect, meaning that test takers are not performing consistently across tasks (defined broadly as including anything from a response to a multiple-choice grammar item to performance on a complex language activity). As testers, we sometime write this variability off as measurement noise or error because a strong interaction effect prevents us from ranking our examinees consistently across the measures. That is, we sometimes attribute this source of variability to error, which then affects our estimate of reliability. More importantly, however, evidence of a test method effect is difficult to interpret. Undoubtedly, a method/interaction effect is due in some part to error. Yet, the issue of person x task interaction mitigates not only our confidence in the reliability of our measures, but in their validity as well. It may very well be that evidence of test method effect is due to our inability to specify a homogeneous construct (Fiske 2002). If our construct definition is not concise and tight we should not be surprised that diverse assessment tasks yield an inconsistent/variable picture of examinee performance. With this statement we turn our attention to how the L2 construct has been defined over the decades. Of particular interest here is to what extent L2 models of language proficiency represent a clearly defined construct, and what the relationship is between a person's language ability and task features (Chapelle 1998).

Chalhoub-Deville and Deville (2003), by examining texts on language testing published since Lado's (1961) seminal work, traced the development in the field over the years of the construct definition of language proficiency. After Lado, some of the most influential works have been (in chronological order): Oller (1979), Canale and Swain (1980), Omaggio (1986), Bachman (1990) and Bachman and Palmer (1996), and McNamara (1996). Chalhoub-Deville and Deville argue that the construct has largely been defined according to a psycholinguistic and cognitive paradigm. Language proficiency has been viewed as an entity that resides in the heads of our test takers (or students or research sample) and is deemed homogenous and static, permitting us to capture a measure of it. Only recently, in the Bachman model, has task been included as an important aspect of the construct representation of language use. (While Oller did promulgate the use of integrated tasks to tap into test takers' expectancy grammars, he never addressed variability in performance.) Although task is

given some prominence in the Bachman model, it is separated from the abilities underlying language use. The separation is maintained to allow generalization of scores based on transferable abilities (see Parkes 2001).

In conclusion, similar to the conceptualization of construct by measurement practitioners, we language testers have made the same assumptions about the L2 construct. The L2 construct is viewed as a stable and homogenous set of ability components while the task is represented as an independent or separate entity. Moreover, performance variability due to interaction between test takers and tasks is seen as a threat to our desire to generalize.

While recognizing the value of a psycholinguistic and cognitive view of language proficiency, we maintain that the consistent finding in the literature of person x task interaction compels us to reconsider the assumption of a homogeneous construct as well as the role of tasks in the representation of the construct. We need to explore other representations of the construct that accommodate findings in the field. Chalhoub-Deville and Deville (2003) advocate considering a perspective based on the role of the social interaction and co-construction of language and knowledge (Kramsch 1986; Johnson 2001; Swain 2001; Young 2000). While some might dismiss this perspective as simply a different wrinkle to mainstream cognitive models of language ability, Chalhoub-Deville (2003) makes a strong case that an "ability-in-language user-in-context" approach (as she calls it) to defining the language construct represents a fundamentally different view of language and has important implications for how we test for and research language ability.

Chalhoub-Deville (1997, 2003) has, for some time, urged language testers to consider context when doing validation work. She maintains that local theories or frameworks are more informative and useful than comprehensive, catch-all models for developing language tests and for theorizing about the construct. Recently she has incorporated thinking from other fields (e.g., aptitude theory) to strengthen and push her argument for the importance of context whenever one investigates the language construct. For her, the construct does not reside in the head of the test taker, but is bound inextricably to the interaction of the person and the task/context. Chalhoub-Deville (2003) concludes her paper by calling for a theory of context.

Variability and context

While the idea of a context dependent representation of the construct might be new to the field of language testing (see, however, Douglas 2002), it has been discussed in the discipline of psychology for some time now (e.g., see Valsiner 1984, and the two edited works, *Toward a Psychology of Situations*, 1981, and

Mind in Context, 1994). Magnusson (1981, 2002), whose work in Sweden in developmental psychology has fostered increased attention to contextual or environmental variables, writes:

> Situations present, at different levels of specification, the information we handle, and they offer us the necessary feedback for building valid conceptions of the outer world as a basis for valid predictions about what will happen and what will be the outcome of our own behaviors. By assimilating new knowledge and new experiences in existing categories and by accommodating old categories and forming new ones, each individual develops a total, integrated system of mental structures and contents in a continuous interaction with the physical, social, and cultural environments. (1981:9)

Magnusson goes on to point out that behaviors take place in certain situations and that it is impossible to understand, explain, predict, or model behavior in isolation from the conditions present in the context.

It is noteworthy that the two gurus of validity theory, Cronbach (1971, 1989) and Messick (1989) seem to disagree in their respective views of context (Moss 1992). Messick seems to hold the view that context is not meaningful for score interpretation, only for generalizing to like contexts. He writes: "the intrusion of context raises issues more of generalizability than of interpretive validity. It gives testimony more to the ubiquity of interactions than to the fragility of score meaning" (p. 15). Messick all but relegates context to a nuisance factor.

In contrast, Cronbach (1989) maintains that the issue of the generalizability of context is tied to construct validation: "Any interpretation invokes constructs if it reaches beyond the specific, local, concrete situation that was observed" (p. 151). Cronbach unambiguously links generalizations across contexts to construct interpretation, something which echoes what was already said with regard to Chalhoub-Deville's (2003) insistence that context can not be separated from construct. A model or framework that treats "ability-in-language user-in-context" as the unit of analysis is called for in both practical and theoretical validation work.

If we are to treat ability-in-language user-in-context as the subject of investigation, we need to be clear as to what we mean by this interaction. Snow (1994), in discussing the relations between person abilities and tasks/situations, lists four interpretations and research approaches to interaction. The first he calls *independent* interaction, which interprets person and task variables independently from each other. Person and situation variables are measured separately and can be interpreted apart from one another. An example from

language testing would be the ILR/ACTFL proficiency scales, where tasks, e.g., reading passages, are selected and administered to specific-ability test takers based on their supposed complexity.

The *interdependent* perspective also sees person and task variables separately, but they exist in relation to one another, e.g., task difficulty must be interpreted with respect to person ability. Bachman (2002) is a strong advocate of this viewpoint, stating "difficulty does not reside in the task alone, but is relative to any given test-taker" (p. 462). Bachman makes this statement in arguing the fruitlessness of studying item/task difficulty factors independently from test takers, yet he stops short of making person-situation the unit of analysis. As stated above, Bachman advocates an approach that clearly distinguishes the two entities from one another.

Snow's next two depictions of interaction differ from the first two in that person and task are now seen as inseparable. The third type of interaction is labeled *reciprocal*, and is characterized by person and task variables acting together to change one another over time. The example Snow provides is a person working on a task who changes or adopts new strategies to solve a problem. Reciprocal interaction would find supporters from the interactionalists and co-constructionists (Kramsch 1986; Johnson 2001; Swain 2001; Young 2000) mentioned above. To illustrate, suppose two test takers were given the task of finding an apartment agreeable to both of them based on descriptions of apartments in a local newspaper and specific likes and dislikes given each test taker. While the tester may have constructed the task with a particular outcome in mind, the test takers may negotiate, compromise, express additional desired attributes of the apartment, and even disagree in the end. The test takers may digress and discuss the pros and cons of city versus rural living. Thus the challenge becomes how to interpret performances related to the intended construct and the unintended – but authentic and related – test behavior. In addition, testers must account for the interactive nature of the task, whereby the performances of the individual examinees are intertwined and the evaluation criteria may have to be flexible.

Snow refers to the fourth approach as *transaction*, where person and situation are in constant reciprocal interaction. In such a system there is no cause and effect, and person-situation constructs exist only in the relations and actions between them. The focus of any scientific inquiry then, from this standpoint, looks at the relations or actions of person-in-situation systems. Again, the interactionalists and co-constructionists might be classified here. A viewpoint that sees the assessment of language for specific purposes as strictly

context bound, yet dynamic and fluid within the particular context, would be transactional.

Snow classifies his own work in aptitude theory as representing both reciprocal interaction and transaction. Important to note is Snow's caution against taking the extreme version of transaction, which he calls "new situationism." This perspective claims that we can learn nothing by studying either person or situation variables, but only their relationship. In arguing against new situationism, Snow cites Allport, who in 1955, wrote:

> … there are *some* things about both organism and environment that can be studied and known about in advance of… [as well as after] their interrelationship, even though the fact that in their relationship they contribute something *to each other* is undeniable. What then, *is* this residual nature or property of the parties to a transaction? (p. 287, emphasis in original)

Snow concludes with a call for eclecticism, saying that each approach to person-situation interaction has some merit.

If ability-in-language user-in-context is to be the unit of analysis when studying language learning and testing, however, we should contemplate how to incorporate the reciprocal interaction and transaction approaches into our thinking and research. One obvious implication is that we consider focusing our research less on the effects of particular variables across individuals and more on the identification, description, and understanding of dynamic systems and processes within person x situation interactions. What aspects of the ability-in-language user-in-context are relevant, which are related, how strong are the relations, are they malleable, how do they change over time? These questions require us to look for patterns within persons, bringing us back to a need for the study of person x task variability mentioned at the outset of this section on language testers as researchers.

According to Magnusson (1981), we must make a distinction between *general* and *differential* effects of situations on behavior, i.e., for our purpose, the effects of context/task on language use. A general effect is the same across individuals, meaning that different contexts affect all test takers similarly, altering their behavior or language use in like ways. Differential effects, however, mean that:

> [a]ccording to an interactional model the characteristic of an individual is in his/her unique *pattern* of stable and changing behaviors across situations of different character or, in terms of data, in his/her partly unique *cross-situational* profile for each of a number of relevant behaviors.
>
> (p. 11, italics in original)

The language test researcher who analyzes the main, general effects of context variables across test takers and dismisses the differential effects of context as unwanted noise or error is assuming a stable, unchanging construct of language ability. Yet, as Chalhoub-Deville and Tarone (1996) argue:

> . . . the nature of the language proficiency construct is not constant; different linguistic, functional, and creative proficiency components emerge when we investigate the proficiency construct in different contexts. (p. 5)

To conclude, we maintain that in our investigations as language test researchers we need to explore constructs not simply to see if they are sensitive to features of the context and task, but also whether their make-up changes dynamically as a communicative situation unfolds. So, in essence, we likely are looking at person x context x process interactions. Trying to unravel the enormous complexity of social – and statistical – interactions brings us full circle back to reliability and validity, i.e., to the question of what our test scores really mean.

Summary and conclusion

In this paper we have reviewed some of the basic concepts of conventional reliability theory and practice and attempted to provide a rationale as to why these considerations are important when test scores are used to make specific decisions about examinees. The traditional views of variability and error, and how these contribute to our ability to generalize our findings and make warranted inferences, are closely tied to the demand for consistency when evaluating test takers. In addition, we delineated some considerations with respect to how language tests are constructed – the kinds of items/tasks used, how these are interrelated, how these are scored – and what implications test construction has for determining variability in test scores, estimating reliability, and marshalling appropriate validity evidence.

Donning the cap of language test researchers we provided some alternative ways of thinking about the basic concepts of reliability, i.e., score variability, and suggested how these might be adopted and adapted in pursuit of an expanded understanding of our construct. We call for the study of ability-in-language user-in-context, where the interaction of language user and context is the focus of study. In addition, we advocate increased attention to the construct of context, and suggest that person x task investigations might offer promise for examining the L2 construct.

Acknowledgements

The authors would like to thank Yong-Won Lee and Carol Chapelle for their insightful and helpful comments and suggestions. Any and all shortcomings to the ideas and arguments presented here, however, must be attributed to the authors.

References

Allport, F. H. (1955). *Theories of perception and the concept of structure*. New York: Wiley.

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: OUP.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*, 453–476.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford: OUP.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, *5*, 225–264.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag New York, Inc.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Chalhoub-Deville, M. (1995a). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, *12*, 16–33.

Chalhoub-Deville, M. (1995b). A contextualized approach to describing oral language proficiency. *Language Learning*, *45*, 251–281.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing*, *14*, 3–22.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*, 369–383.

Chalhoub-Deville, M. & Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (Ed.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 815–832). Harlow, England: Longman.

Chalhoub-Deville, M. & Tarone, E. (1996). What is the role of specific contexts in second-language acquisition, teaching, and testing? Unpublished manuscript.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Second language acquisition and language testing interfaces* (pp. 32–70). Cambridge: CUP.

Chapelle, C. A. (1999). Validity in language assessment. In W. Grabe (Ed.), *Annual Review of Applied Linguistics* (pp. 254–272). Cambridge: CUP.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement theory and public policy* (pp. 147–171). Urbana, IL: University of Illinois Press.

Deville, C. & Chalhoub-Deville, M. (1993). Modified scoring, traditional item analysis and Sato's caution index used to investigate the reading recall protocol. *Language Testing*, *10*, 117–132.

Douglas, D. (2002). *Assessing language for specific purposes.* Cambridge: CUP.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: OUP.

Feldt, L. S. (2002). Estimating internal consistency reliability of a tests composed of testlets varying in length. *Applied Measurement in Education*, *15*, 33–48.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education and Macmillan.

Feldt, L. S. & Charter, R. A. (2003). Estimating the reliability of a test split into two parts of equal or unequal length. *Psychological Methods*, *8*, 102–109.

Fiske, D. W. (2002). Validity for what? In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 169–178). Mahwah, NJ: Lawrence Erlbaum.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*, 523–561.

Johnson, M. (2001). *The art of nonconversation: A re-examination of the validity of the oral proficiency interview*. New Haven, CT: Yale University Press.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*, 366–372.

Lado, R. L. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book*. New York: McGraw-Hill.

Lee, G. (2002). The influence of several factors on reliability for complex reading comprehension tests. *Journal of Educational Measurement*, *39*, 149–164.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice*, *19*, 9–15.

Lee, G., Dunbar, S. B., & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing test composed of testlets. *Educational and Psychological Measurement*, *61*, 958–975.

Lee, Y-W., Kantor, R., & Mollaun, P. (April, 2002). Score dependability of the writing and speaking sections of New TOEFL. Paper presented at the annual meeting of the *National Council on Measurement in Education* (*NCME*), New Orleans, LA.

Linn, R. L. & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*(1), 5–8, 15.

Magnusson, D. (Ed.). (1981). *Toward a psychology of situations.* Hillsdale, NJ: Lawrence Erlbaum.

Magnusson, D. (1981). Wanted: A psychology of situations. In D. Magnusson (Ed.), *Toward a psychology of situations: An interactional perspective* (pp. 9–35). Hillsdale, NJ: Lawrence Erlbaum.

Magnusson, D. (2002). The individual as the organizing principle in psychological research: A holistic approach. In L. R. Bergman, R. B. Cairns, L.-G. Nilsson, & L. Nystedt (Eds.), *Developmental science and the holistic approach* (pp. 33–47). Mahwah, NJ: Lawrence Erlbaum.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, *18*, 333–349.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*, 229–258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5–12.

Nesselroade, J. R. & Ghisletta, P. (2000). Beyond static concepts in modeling behavior. In L. R. Bergman, R. B. Cairns, L.-G. Nilsson, & L. Nystedt (Eds.), *Developmental science and the holistic approach* (pp. 121–135). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Oller, J. W., Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Omaggio, A. C. (1986). *Teaching language in context: Proficiency-oriented instruction*. Boston, MA: Heinle & Heinle Publishers.

Parkes, J. (2001). The role of transfer in the variability of performance assessment scores. *Educational Assessment*, *7*, 143–164.

Qualls, A. L. (1995). Estimating reliability of a test containing multiple item formats. *Applied Measurement in Education*, *8*, 111–120.

Shavelson, R. J., Baxter, G. P., & Goa, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215–232.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*, 99–123.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.), *Mind in context* (pp. 3–37). Cambridge: CUP.

Sternberg, R. J. & Wagner, R. K. (Eds.). (1994). *Mind in context*. Cambridge: CUP.

Swain, M. (1993). Second language testing and second language acquisition: Is there a conflict with traditional psychometrics? *Language Testing*, *10*, 193–207.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, *18*, 275–302.

Tarone, E. (1988). *Variation in interlanguage*. London: Edward Arnold.

Valsiner, J. (1984). Two alternative epistemological frameworks in psychology: The typological and variational modes of thinking. *The Journal of Mind and Behavior*, *5*, 449–470.

Young, R. F. (March, 2000). Interactional competence: Challenges for validity. Paper presented at the annual meeting of the *Language Testing Research Colloquium* (*LTRC*), Vancouver, Canada.

Yu, C. H. (April, 2003). Misconceived relationships between logical positivism and quantitative research. Paper presented at the annual meeting of the *American Educational Research Association* (*AERA*), Chicago, IL.

Validity and values

Inferences and generalizability in language testing

Tim McNamara
The University of Melbourne

Language testing research is an increasingly divided field, as it responds to the paradigm shifts in broader applied linguistics research. On the one hand, language testing validation research places a fundamental emphasis on the generalizability of results and the appropriateness of inferences made based on observed learner performances. This involves a rigorous interrogation of the elicitation instruments, judgments, and observations used to make inferences about individual test takers. At the same time, input from non-measurement traditions is leading to the exploration of new insights into the limitations of such inferences, and to a greater understanding of the social values which imbue tests. This epistemological ferment is as much productive as problematic.

Introduction

In this chapter, I want to review recent work within language testing about the issue of generalizing from learner performances, that is, drawing inferences about the test-taker's ability beyond the immediate testing context. Of all the fields of applied linguistics, it is in language testing that questions of inferences and generalizability are arguably of most fundamental concern. Language testing researchers (e.g. Bachman 1990; Bachman 2005; Chapelle 1998; Chapelle et al. 2004; McNamara 2003), drawing on work in educational measurement (especially Messick 1989; Mislevy 1996; Kane 1992, 2004), have long argued that test scores represent inferences or generalizations about individual test-takers, and discuss procedures for investigating their validity. There are two related aspects of these discussions of generalizability from test performance which I would like to address in this chapter. The first is the way in which we may assemble reasoned arguments and empirical evidence in support of the

validity of inferences from test scores. The second is about the challenge to the procedures typically used for gathering such evidence presented by critiques of positivism within the social sciences and applied linguistics. These critiques require us to consider to the values implied within test constructs, and the social and political context of the uses made of test scores. In this latter domain, the discussions within language testing mirror the broader 'paradigm debate' about the epistemology of research in other fields of applied linguistics.

Drawing generalizable inferences in language assessment

Language tests are procedures for generalizing. Scores on language tests are reports of performance in the test situation, but they only have meaning as inferences *beyond* that situation. The basic logic of generalizing from observed performance under test conditions to performance under non-test conditions, which is the real target of the assessment, can be set out as in Figure 1.

The first step is to identify the assessment target – the inferences we would like to draw about learners: what they know, what they can do, what performances they are capable of *beyond* the assessment setting. This target of test inferences is known as the *criterion domain*. This domain cannot be 'known' in any direct sense, but only surmised indirectly, 'through a glass, darkly', in an assessment version of Labov's Observer's Paradox, or Heisenberg's Uncertainty Principle.

Now, as daily experience proves, inferences and surmises *can* be rather wide of the mark. In order to constrain test inferences, careful attention is paid to

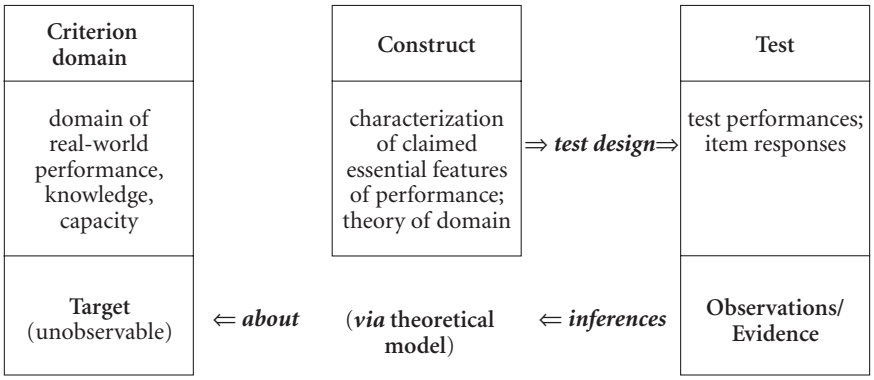

Figure 1. Criterion, construct and test

thinking through exactly what it is that we want to infer, and developing *procedures* for gathering the information that forms the basis for test inferences. This involves two stages.

First, the criterion domain is modelled; this is the assessment *construct*. The construct is always in a sense provisional, and is subject to ongoing revision, as understanding of the domain, the nature of language, language use, and of learning, changes over time. The source of domain modelling for the purpose of assessment may be the curriculum, which itself constitutes a model, a view of the target domain, its structure and principal divisions and characteristics, and assessments may accordingly draw on curriculum constructs as the basis for interpreting the data of learner performance. But there are other sources, in psycholinguistics, in sociolinguistics, in discourse theory. Such constructs embody social, educational and political values, and are never neutral; they are constantly subject to intellectual contestation, and to social and political influence. I will say more about constructs in a moment.

Second, the construct informs the development of a procedure for principled observation of performance under known conditions. This is the *assessment* or, if it is done formally, the *test*. Ideally, the tester is now in a position to draw inferences on the basis of these observations either about the probable character of performance under non-test conditions, or about the candidate's standing in relation to a domain of knowledge and abilities of interest.

In summary, then, crucial to assessment is the distinction between the *criterion domain* (the target of test inferences) and the *test* (the evidence from which inferences are drawn). And because these inferences are necessarily indirect, they are mediated through test *constructs*, that is, modelling of the criterion domain in terms of what are held to be its essential features or characteristics.

Validating test score inferences

The inferences drawn in language tests, given their inevitable uncertainty, are open to challenge. Much research in language testing is motivated by a scepticism in principle about the meaningfulness and defensibility of the inferences we make based on test performance.[1] This notion is the crucial feature of the classic validity framework of Messick (1989) (Figure 2), and of subsequent developments of it in the work of Mislevy (1996), Kane (1992, 2004) and others.

Messick distinguishes a number of facets of validity within a unified theory of validity. The first cell in the matrix emphasises the need to use *reasoning*

	Test interpretation	Test use
Evidential basis	Construct validity	Construct validity + Relevance / utility
Consequential basis	Value implications	Social consequences

Figure 2. Facets of validity (Messick 1989: 20)

and *empirical evidence* in support of the constructs on which test score interpretations are based. The second cell on the top line stresses the point that such inferences are never context free, and need to be considered in the context in which the assessment is to be used. Test score inferences need to be revalidated for every major context of use. For example, this should make us question the current use of scores on IELTS (a test of English for academic and general training purposes) in order to determine language proficiency in the context of immigration selection in Australia and other countries, a use for which IELTS has not to date been validated. The third and fourth cells take us into the realm of the values implicit in test construct, and the social consequences of test use; we will discuss these below.

Constructs and evidence as the basis for test score inferences

As stated above, *reasoning* about the construct is the first stage of test development and the basis of much critique of the validity of the inferences that are drawn from test performance. Usually, this reasoning draws on theories of language knowledge and language performance which are essentially cognitive, so that the inferences drawn are about the cognitive characteristics of test takers. But constructs can also be social in character, so that test performance may be the basis for drawing inferences about the social characteristics of test takers – their social identity. This is an important function of language tests, familiar since the shibboleth test of the Bible. Let me give you a recent example. For the past 15 years, and increasingly, the claims of those seeking refugee status in a number of countries including the U.K., many other European countries, Australia and New Zealand, have been determined in part on the basis of a language test. For example, one large group of refugee claimants recently arriving in Australia and other countries is composed of members of the Hazara minority from Afghanistan, who have been subject to persecution in Afghanistan on religious and ethnic grounds for many years. Hazaras speak a dialect of Dari, one of the two main languages of Afghanistan, itself close to

the Farsi spoken in Iran. The applications of these claimants are routinely subject to query on the grounds that the individuals making the claims may not be from Afghanistan but from established Hazara communities in neighbouring countries such as Pakistan or Iran, where they are not subject to persecution and thus under international law have no right to refugee status when they arrive in another country. The refugees are interviewed by an immigration officer through an interpreter, and the tape of the interview is sent to so-called language experts employed by private companies (as it happens, based mainly in Sweden). On the basis of the language evidence on the tape, the 'experts' draw their inferences: they claim to be able to tell, based on features of lexical choice and pronunciation, whether the person is a Hazara from inside Afghanistan, or from Pakistan or Iran. Leaving aside for a moment the competence and the methods of the 'experts' involved, the validity of these inferences is rendered questionable by the problematic nature of the construct underlying the test. The issue is that the boundaries of the linguistic community of speakers of the Hazara dialect do not coincide with the political border between Afghanistan and Pakistan, and there is simply not the sociolinguistic information available to determine the issue accurately. Obviously, given the conditions in Afghanistan over the last thirty years, there has been no proper sociolinguistic work on the varieties concerned and their geographical distribution; and the disruption caused by the years of war, the refugee flows, the influence of teachers who are speakers of other varieties, and so on mean that sociolinguistically the situation is likely to be very fluid. We can thus object to the procedure, and to the quality of the inferences to which it leads, by arguing about the construct, as a group of sociolinguists, phoneticians and language testers have recently done (Eades et al. 2003).

Most language testing research is about the *evidential basis* for test score interpretation and use. In the case of the asylum seekers, for example, empirical evidence of the extent of false negative and false positive identifications has been gathered in the form of data on asylum seekers who have been rejected by countries to which they have been sent 'home' – 20% of cases, in one Swedish report. This empirical evidence thus confirms the earlier theoretical doubts on the validity of inferences drawn on the basis of the procedure. More typically, empirical evidence is sought from analysis of responses by candidates to test items. Does the evidence from these responses support the structural relationships predicated in the construct? Particularly helpful here is Messick's notion of construct-irrelevant variance. For example, if the chances of a score on an oral proficiency test depend in part on the luck of the draw in terms of which interviewer interacts with the candidate (Brown 2003), then part of

the score variance will reflect characteristics of the interlocutor rather than the candidate, with subsequent threats to the generalizability of inferences about candidates. An alternative way of understanding this issue is to see it as involving construct definition. Candidate performance in this case involves joint action on the part of both candidate and interviewer, and cannot therefore be seen as a simple projection of the individual candidate's ability (McNamara 1997; Chalhoub-Deville 2003).

Note that this investigation of the relationship between test construct and test performance cuts both ways: test data can not only be used to validate, in other words to question, the inferences that we draw based on test constructs, but can also be used to question the constructs on which our inferences have been based. In this way, just as SLA is a source of constructs for language testing, language testing is a site for research in second language acquisition. The study reported in Iwashita et al. (2001) is a case in point. The research involved attempting to build a task-based speaking test drawing on the framework proposed by Skehan (1998). The study demonstrated not only that it was not possible to build a test of speaking that would yield valid scores by using the framework in this context, but that the test data raised interesting questions about the model itself. In other words, in the attempt to validate score inferences from the test, validity evidence relevant to the construct itself was gathered.

Recently, other validity theorists have operationalized the aspects of Messick's validity framework discussed so far. This work has resulted in two accomplishments. First, in the work of Mislevy (1996), it stresses the logical steps necessary for validation work, and offers conceptual clarification of the design of validity studies. Second, in the work of Kane (1992, 2004), it establishes a more manageable process of validation, made necessary by the fact that the very complexity of the validation issues raised by Messick threatens to overwhelm test developers.

Mislevy (1996; Mislevy et al. 2002, 2003) proposes what he calls evidence centred test design, in which he distinguishes the claims we wish to make about test-takers (the inferences we wish to be able to draw about them), what we would need to be able to observe about test takers in order to form a basis for drawing those inferences, and the conditions under which this evidence might be sought. The logical analysis so entailed of claims, evidence and task design leads then to the design of empirical procedures for gathering data from test performance and analysing it statistically in order ultimately to validate the claims.

Kane (1992, 2004; Kane et al. 1999) proposes the need to develop a validation argument, with carefully defined stages involving procedures for generalizing from the evidence of test performance to the inferences one wishes to draw about candidates, and for envisioning the main threats to the validity of those inferences in order to design appropriate studies to investigate such threats.

Generalization as reliability

The centrality of generalizability in language testing is underlined by the fact that we are not interested in the test performance for its own sake: it has no meaning or value in and of itself. It has meaning only as a basis for generalizing beyond the particular performance. Raters scoring an essay written for a language test are rarely interested in the content of the essay as such; this written communication between an anonymous candidate and an anonymous rater is a performance for a further purpose. It is a deliberate display of abilities on the part of the candidate in accordance with the rules for display set up by the tester so that the abilities can be assessed. In a speaking test, the interlocutor is not necessarily interested in what the candidate has to say in itself; or if they do happen to find it personally interesting, this is at best irrelevant, and at worst a source of construct-irrelevant variance! But there is one further specific issue of generalizability which is routinely addressed in the analysis of test scores, and that is the question of the repeatability of the inference drawn. This generally involves the need to establish the *reliability* of the procedure. In tests involving a series of discrete items, as commonly found in listening, reading, grammar or vocabulary tests, the response to each item can be considered as a separate performance, and investigation of reliability involves checking to see whether the inferences about candidates drawn from performances on individual items or sets of items support one other, that is, whether items ostensibly testing similar things yield similar inferences about the candidate. For tests involving sets of items, such as most grammar, vocabulary, listening or reading tests, this is done technically by looking at the discrimination of each item, and by aggregating item discrimination indices as a measure of overall reliability. In tests of productive performance such as tests of speaking and writing, scoring involves human judgement, and issues of repeatability here involve estimating the extent of agreement between judges – in other words, would the candidate get the same score next time with a different judge?[2]

Epistemology and research methodology in investigating the validity of test score inferences

Addressing the evidence-related issues outlined above typically involves statistical methods, which have become increasingly sophisticated; a helpful review of these developments is available in Bachman (2000). They include Generalizability Theory, Item Response Modelling, including Rasch modelling, Structural Equation Modelling and others. But the use of such psychometric techniques in language testing invites charges that language testing research is irredeemably positivist, and thereby unable to respond to the challenge of the 'paradigm debate' about the epistemology of research in the wider arena of the social sciences. The epistemological assumptions of positivism and post-positivism, critiques of them, and alternative research paradigms, are helpfully set out by Brian Lynch in his book *Language Program Evaluation* (Lynch 1996). Lynch frames this debate in part in terms of competing notions of the validity of inferences in research, and in particular whether, and if so how, we are to generalize from research data. Clearly, the very themes of this volume – issues of inferencing and generalizability – have been triggered by and are reflective of this wider debate. Within applied linguistics this debate has been evident in heated, even acrimonious exchanges over conflicting epistemologies for carrying out research on Second Language Acquisition (Block 1996; Gregg et al. 1997; Firth & Wagner 1997).

Research on language testing has for some time similarly reflected this paradigm debate. Brian Lynch and Liz Hamp-Lyons (Lynch & Hamp-Lyons 1999; Hamp-Lyons & Lynch 1998) have examined the epistemological bases for research papers given at the Language Testing Research Colloquium over a number of years, and detected a trend away from hard positivism, although the field is still one of the most positivist in orientation of all areas of Applied Linguistics. One common way of interpreting the issues in the debate is in terms of competing research methodologies (quantitative or qualitative). A growing preference for qualitative research methods in applied linguistics is reflected in the changes over the years in the relative proportions of qualitatively and quantitatively based research papers in applied linguistics journals, as tracked by Lazaraton (2000) and others. This trend is also evident within language testing, as Lynch and Hamp-Lyons show (see also Lumley & Brown 2004).

The use of qualitative methodologies is found in many test validation contexts, for example the use of introspective methods in research on multimedia listening and viewing (Gruba 1999), listening (Buck 1991; Yi'an 1998), the rating of writing (Lumley 2000), and many others. A recent survey of discourse

Table 1. Qualitative research on the validation of oral tests

Research methods used	Topic	Aspect
Activity Theory Conversation Analysis	Oral Proficiency Interview	Institutional character of event
Conversation Analysis Discourse Analysis	Interlocutor effects in traditional proficiency interviews	Characterisation of interlocutor skill Impact of gender Impact of familiarity of interlocutor Impact of native speaker status of interlocutor
Discourse Analysis Activity Theory	Interlocutor effects in paired and group oral tests	Impact of personality Impact of proficiency
Discourse Analysis Interviews/retrospective recall	Effect of Task	Role play versus interview 'Authentic' interactions with native speakers Monologue vs dialogue Functions elicited Information gap tasks Task performance conditions (Skehan)
Conversation Analysis Discourse Analysis	Test taker characteristics	Impact of gender Impact of cultural background
Think aloud Discourse Analysis	Rating scales and rater cognition	Interpretation of criteria Evolution of criteria Development of rating scales
Discourse Analysis	Classroom assessment	Formative assessment
Discourse Analysis Case studies	The use of technology in speaking tests	Semi-direct tests of speaking (SOPI)

based studies of oral language assessment (McNamara, Hill, & May 2002) identified the extent of qualitative research methods used in such research. Table 1 sets out the qualitative research methods used and the topics dealt with in these studies. Reference details for the research listed here are available in McNamara, Hill and May (2002).

As well, significant studies of the validity of oral language tests have used a combination of quantitative and qualitative methods. For example, O'Loughlin (2001) combines Rasch analysis, lexical density counts of candidate discourse, and case studies of individual candidates to establish his claim about the lack of comparability of two supposedly equivalent instruments, direct and semi-direct measures of oral proficiency. Iwashita et al. (2001) use a combination of Rasch analysis and careful study of candidate discourse to reach important conclusions about the applicability of Skehan's model of task performance conditions to the definition of dimensions of task difficulty in a speaking test.

Brown (2003) combines Conversation Analysis and Rasch analysis of scores to illuminate the character of interlocutor behaviour in speaking tests and its impact on scores.

The choice of qualitative or quantitative research methods in validation studies of this kind does not appear then to involve a difficult epistemological choice. It is fair to characterize all such research aiming at providing evidence for or against the legitimacy of inferences from test scores, both quantitative and qualitative, as empiricist. This commonality is I think more significant than any differences or preference for qualitative over quantitative methods on *a priori* epistemological grounds.

Values and consequences in language tests

A more profound epistemological challenge within language testing research is associated with a deepening awareness of the social and political values embodied both in language constructs and in language testing practice. Let us return to Messick's validity matrix (Figure 2, above).

The bottom row of the matrix insists that all test constructs, and hence all interpretations of test scores, involve questions of value. Messick in other words sees all testing as fundamentally political in character – by that I mean that its constructs and practices are located in the arena of contestable values. This requires us to reveal and to defend the value implications of our test constructs (cell 3), and to consider the impact of tests, and in particular, the wanted and unwanted consequences of interpretations about test-takers in contexts of test use (cell 4). Messick (1989) locates these requirements in an extensive discussion of the epistemology of measurement research and practice, and demonstrates a deep awareness of the critiques of positivism. Moss (1998) similarly locates her concern for test consequences in a discussion of the work of Foucault, Bourdieu and other critics of the positivism of the social sciences.

The educational measurement field, like language testing, responded strongly to the idea that exploration of the consequences of the use of language test scores be seen as part of test score validation. While many figures endorsed this position (Shephard 1997; Linn 1997; Kane 2001), pointing out that validation of the *use* of test scores had been a central theme of the discussion of validity theory for over 40 years, others (Popham 1997; Mehrens 1997) objected on the grounds that it was not helpful to see this as part of validation research, which they wanted to restrict to investigation of the accuracy of test score interpretation. In some ways, however, this is a dispute more over

wording than substance, as both Popham and Mehrens agree that the consequences of test use *are* a subject requiring investigation; they simply do not want to call it 'validation' research. In fact, Popham has recently published a book (Popham 2001) on what he sees as the damaging consequences of school and college testing in the United States.

The response of the field of language testing to Messick's notion of the value-laden character of assessment was slow at first. This was in part due to the fact that this aspect of the work was relatively underplayed in the influential original presentation of Messick's ideas in the work of Bachman (1990), who perhaps understandably emphasized the important implications for language testing research of the top two cells of the matrix.

In terms of the *values* implied in language test constructs, there is a growing critique of individualistic models of proficiency (McNamara 1997; Chalhoub-Deville & Deville 2004). A strong influence here has been work on discourse analysis which stresses the co-construction of performance (Jacoby & Ochs 1995: 177):

> One of the important implications for taking the position that everything is co-constructed through interaction is that it follows that there is a distributed responsibility among interlocutors for the creation of sequential coherence, identities, meaning, and events. This means that language, discourse and their effects cannot be considered deterministically preordained … by assumed constructs of individual competence…

This work has been most advanced in the area of oral language proficiency testing in face-to-face contexts, such as in the oral proficiency interview (e.g. Brown 2003) or paired-candidate interactions (e.g. Iwashita 1998). In each case, the performance of the interlocutor (interviewer, other candidate) as implicated in the performance of the candidate poses difficult questions for assessment schemes which are framed in terms of reports on individual competence. In the words of Schegloff (1995: 192):

> It is some 15 years now since Charles Goodwin … gave a convincing demonstration of how the final form of a sentence in ordinary conversation had to be understood as an interactional product… Goodwin's account … serves… as a compelling call for the inclusion of the hearer in what were purported to be the speaker's processes.

Recent work has drawn attention to the potential of poststructuralist thought in understanding how apparently neutral language proficiency constructs are inevitably socially constructed and thus embody values and ideologies (McNamara 2001, 2006). It is worth noting here that the deconstruction of

such test constructs applies no less to constructs in other fields of applied linguistics, notably second language acquisition.

There is also a growing realization that many language test constructs are explicitly political in character and hence not amenable to influences which are not political (Brindley 1998, 2001; McNamara 2006). Examples are the constructs embedded in competency based language assessments used in adult vocational education and training as a direct result of government policy. In Australia, the ESL assessment of adult immigrants via the Certificate in Spoken and Written English (CSWE: Burrows 1993), which ostensibly is based loosely on a Hallidayan model of language use, is better understood as an expression of the outcomes based and functionalist demands of current government policy on adult education and training. A more recent example is the extraordinary role within Europe (and beyond) of the Common European Framework for Languages (Council of Europe, 2001). The political imposition of language testing constructs means that language test validation research will have an impact only to the extent that it is politicized. This is at odds with the liberal notion that academic research on language test validity will have an impact in and of itself, that is, through the sheer power of its reasoning. Thus it is possible that what is most significant about the Code of Ethics of the International Language Testing Association (ILTA) (ILTA, 2000) is its political character, the power it may have institutionally to affect practice, rather than its inherent moral persuasiveness. ILTA itself is a potentially political force, although this potential is far from being maximized at present.

The relative impotence of scholars is not of course a new story. In a discussion of the role of Confucian thought in the Han Dynasty in China in the 3rd century BC, the late eminent Harvard historian John King Fairbank writes (Fairbank & Goldman 1998: 63):

> As W. T. de Bary (1991) points out, the Confucians did not try to establish 'any power base of their own… they faced the state, and whoever controlled it in the imperial court, as individual scholars… this institutional weakness, highly dependent condition, and extreme insecurity…. Marked the Confucians as *ju* ('softies') in the politics of imperial China.' They had to find patrons who could protect them. It was not easy to have an independent voice separate from the imperial establishment.

These explicitly political considerations bring us to the implications of the notion of test *consequences*. Returning to the example of the asylum seekers given earlier in this paper, a concern for language test consequences means that even were the procedure currently used to be carried out on the basis of

a more adequate and defensible sociolinguistic construct, it necessarily has serious consequences for those concerned, and raises complex issues of ethics. The ethical responsibilities of language testers as an aspect of their professionalism is now a major topic of debate within the profession, for example in the work of Spolsky (1981, 1997), Hamp-Lyons (1997), Kunnan (1997, 2000), and particularly Davies (1997, 2004).

A more radical response to the issue of the consequences of test use involves direct political critique of the institutional character of assessment. Most significant here is the work of Elana Shohamy (1998, 2001a, 2001b; see also Lynch 2001), who, following Pennycook (2001), has introduced the term *critical language testing* to describe a research orientation committed to exposing undemocratic practices in assessment. This important work is still in its early stages. An excellent example of the potential for a kind of Foucauldian archival research on language tests is the account in Spolsky (1995) of the political and institutional forces surrounding the initial development and struggle for control of the Test of English as a Foreign Language (TOEFL).

A growing awareness of the largely managerial and institutional function of language assessment has been accompanied by an increasing sense of the neglect by assessment research of the needs of teachers and learners as they engage with the process of learning. This neglect is evident in the ongoing focus on high stakes proficiency assessment as the main staple of academic research on language testing, for example as measured by papers in *Language Testing* or papers given at the Language Testing Research Colloquium (LTRC), although recent years have seen a welcome correction to this picture, with a special issue of *Language Testing* devoted to this subject, and a strong thematic focus on classroom assessment at the 2004 LTRC. Particularly problematic is the fact that the very notion of generalizability of assessment, so fundamental (as we have seen) to current discussions of validity, can sometimes compromise the potential of assessment to serve the needs of teachers and learners. The need for assessments to be comparable, the basis for reliability, one aspect (as we have seen) of generalizability, places constraints on classroom assessment, particularly in the area of communicative skill (speaking and writing), that most teachers cannot meet (Brindley 2000). Moreover, any attempt to meet such demands constrains the possibilities of assessment in the classroom. We urgently need to explore ways in which assessment concepts can be better used in classrooms to benefit learning and teaching (Leung 2004).

In conclusion, I think that language testing researchers should welcome and participate vigorously in the debates about epistemology and research paradigms as relevant to language testing. Engagement with debate of this kind

is precisely the sort of interrogation encouraged by validity theory as envisaged by Messick in particular. It will be clear that I think that we should not be constrained from embracing the full range of research methods open to us, be they neo-positivist or otherwise, if they deepen (as I believe they do) our ability to understand the bases for inferences in language tests. The greater challenge, I believe, is to use a recognition of the value-laden and political character of language tests to re-think their deeper meaning and social function. But this will not be easy, if the work of some of Messick's heirs is an indication. Mislevy's otherwise brilliant work at no time engages with the issue of the consequences of test use. And while Kane does in principle embrace a concern for consequences, he does not develop a methodology for reflecting on them or investigating them other than through the general structure of the validation argument he proposes, which reduces the issue to a technical and empirical one. This narrow position is reflected in a recent paper of Bachman (2005), who, similarly concerned with the manageability of the validation task, proposes that test use consequences be investigated through a second, parallel validation argument modelled on that used to validate test score interpretation as it were outside a context of use. I would argue that the focus of our concern about the consequences of test score use and the values implied in testing practice needs to be much broader. Here I will end with the words of Moss (1998: 11). In discussing the social impact of assessment practice, she says:

> This perspective on the dialectical relationship between social reality and our representation of it has implications for understanding the crucial role of evidence about consequences in validity research. ... The practices in which we engage help to construct the social reality we study. While this may not be apparent from the administration of any single test, over time the effects accumulate. ... While this argument has some implications for validity theory as it relates to specific interpretations and uses of test scores, the import spills over... to encompass the general practice of testing. ... The scope of the [validity] argument goes well beyond ... test specific evaluation practices; it entails ongoing evaluation of the dialectical relationship between the products and practices of testing, writ large, and the social reality that is recursively represented and transformed.

Notes

1. It must be said that there are schools of thought in language testing which don't always reflect such scepticism: three come to mind:

– the believers in direct testing, such as many of the proponents of the ACTFL scales (Liskin-Gasparro 1984; Lowe 1985), and in general those advocating what Bachman (1990) calls the 'real life' approach;
– many of those promoting a scale and framework approach to language testing, such as the Council of Europe's Common European Framework of Reference for Languages (Council of Europe, 2001) where concerns about empirical validation of the scales in question (North & Schneider 1998) are matched, if not surpassed, by efforts to negotiate their political acceptability (Trim 1997);
– the small but growing influence of systemic functional linguistics in language assessment (Matthiessen, Slade, & Macken 1992; Mincham 1995; South Australian Curriculum and Standards Authority, 2002), which stresses the 'objectivity' of the assessment as a matter of principle rather than of empirical investigation.

2. The demand for reliability proves, however, to be problematic in certain key assessment contexts, particularly the classroom, a point to be addressed near the end of the paper.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: OUP.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1–42.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

Block, D. (1996). Not so fast: Some thoughts on theory culling, relativism, accepted findings and the heart and soul of SLA. *Applied Linguistics, 17*(1), 63–83.

Brindley, G. (1998). Outcomes-based assessment and reporting in language programs: A review of the issues. *Language Testing, 15*, 45–85.

Brindley, G. (2000). Task difficulty and task generalisability in competency-based writing assessment. In G. Brindley (Ed.), *Issues in immigrant English language assessment*. Volume 1 (pp. 45–80). Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing, 18*(4), 393–407.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1–25.

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing, 8*(1), 67–91.

Burrows, C. (1993). *Assessment guidelines for the Certificate in Spoken and Written English*. Sydney: New South Wales Adult Migrant Education Service.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.

Chalhoub-Deville, M. & Deville, C. (2004). A look back at and forward to what language testers measure. In T. McNamara, A. Brown, L. Grove, K. Hill, & N. Iwashita (Section Eds.), Second language testing and assessment, Part VI of E. Hinkel (Ed.), *Handbook Of research in second language teaching and learning*. Mahwah, NJ: Lawrence Erlbaum.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). New York: CUP.

Chapelle, C. A., Enright, M. E., & Jamieson, J. (2004). Issues in developing a TOEFL validity argument. Paper presented at the Language Testing Research Colloquium, Temecula, CA, March.

Council of Europe (2001). *A common European framework of reference for language learning*. Cambridge: CUP.

Davies, A. (Ed.). (1997). Special issue: Ethics in language testing. *Language Testing, 14*(3).

Davies, A. (Ed.). (2004). Special issue: The ethics of language assessment. *Language Assessment Quarterly, 2&3*.

Eades, D., Fraser, H., Siegel, J., McNamara, T., & Baker, B. (2003). Linguistic identification in the determination of nationality: A preliminary report. *Language Policy, 2*(2), 179–199.

Fairbank, J. K. & Goldman, M. (1998). *China: A new history* (enlarged edition). Cambridge, MA and London: Belknap Press of Harvard University Press.

Firth, A. & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *Modern Language Journal, 81*, 285–300.

Gregg, K. R., Long, M. H., Jordan, G., & Beretta, A. (1997). Rationality and its discontents in SLA. *Applied Linguistics, 18*(4), 538–558.

Gruba, P. (1999). The role of digital video media in second language listening comprehension. PhD Dissertation, University of Melbourne.

Hamp-Lyons, L. (1997). Ethics in language testing. In D. Corson (Series Ed.) & C. Clapham (Vol. Ed.), *Language testing and assessment:* Vol. 7, *Encyclopedia of language and education* (pp. 323–333). Dordrecht: Kluwer.

Hamp-Lyons, L. & Lynch, B. K. (1998). Perspectives on validity: A historical analysis of language testing conference abstracts. In A. J. Kunnan (Ed.), *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium, Long Beach* (pp. 253–276). Mahwah, NJ: Lawrence Erlbaum.

International Language Testing Association (ILTA) (2000). Code of ethics for ILTA. *Language Testing Update, 27*, 14–22.

Iwashita, N. (1998). The validity of the paired interview in oral performance assessment. *Melbourne Papers in Language Testing, 5*(2), 51–65.

Iwashita, N., McNamara, T. F., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning, 51*(3), 401–436.

Jacoby, S. & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction, 28*(3), 171–183.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 12*, 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(3), 135–170.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

Kunnan, A. (1997). Connecting fairness and validation. In A. Huhta, V. Kohonen, L. Kurki-Suomo, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 85–105). Jyväskylä, Finland: University of Jyväskylä.

Kunnan A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment. Selected papers of the 19th Language Testing Research Colloquium. Orlando, Florida* (pp. 1–10). Cambridge: UCLES & CUP.

Lazaraton, A. (2000). Current trends in research methodology and statistics in applied linguistics. *TESOL Quarterly*, *34*(1), 175–181.

Leung, C. (2004). Classroom teacher assessment of second language development: Construct as practice. In T. McNamara, A. Brown, L. Grove, K. Hill, & N. Iwashita (Section Eds.), Second language testing and assessment, Part VI of E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*. Mahwah, NJ: Lawrence Erlbaum.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14–16.

Liskin-Gasparro, J. E. (1984). The ACTFL proficiency guidelines: Gateway to testing and curriculum. *Foreign Language Annals, 17*(5), 475–489.

Lowe, P. (1985). The ILR proficiency scale as a synthesising research principle: The view from the mountain. In C. J. James (Ed.), *Foreign language proficiency in the classroom and beyond* (pp. 9–53). Lincolnwood, IL: National Textbook Company,

Lumley, T. (2000). The process of the assessment of writing performance: The rater's perspective. PhD Dissertation, The University of Melbourne.

Lumley, T. & Brown, A. (2004). Research methods in language testing. In T. McNamara, A. Brown, L. Grove, K. Hill, & N. Iwashita (Section Eds.), Second language testing and assessment, Part VI of E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*. Mahwah, NJ: Lawrence Erlbaum.

Lynch, B. K. (1996). *Language program evaluation*. Cambridge: CUP.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18*(4), 351–372.

Lynch, B. K. & Hamp-Lyons, L. (1999). Perspectives on research paradigms and validity: Tales from the Language Testing Research Colloquium. *Melbourne Papers in Language Testing, 8*(1), 57–93.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446–466.

McNamara, T. F. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333–349.

McNamara, T. F. (2006). *Validity in language testing: The challenge of Sam Messick's legacy. Language Assessment Quarterly, 3*(1).

McNamara, T. F., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221–242.

McNamara, T. F. (2003). Validity and reliability in the senior school curriculum: New takes on old questions. Invited presentation, Australasian Curriculum, Assessment and Certification Authorities (ACACA) 2003 National Conference, Adelaide, July.

Matthiessen, C. M. I. M., Slade, D., & Macken, M. (1992). Language in context: A new model for evaluating student writing. *Linguistics in Education, 4*(2), 173–195.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16–18.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Mincham, L. (1995). ESL student needs procedures: An approach to language assessment in primary and secondary school contexts. In G. Brindley (Ed.), *Language assessment in action* (pp. 65–91). Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379–416.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of assessment arguments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6–12.

North, B. & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing 15*(2), 217–263.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests. Studies in Language Testing, 13*. Cambridge: UCLES/CUP.

Pennycook, A. (2001). *Critical applied linguistics: A critical introduction*. Mahwah, NJ: Lawrence Erlbaum.

Popham, W. J. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13.

Popham, W. J. (2001). *The truth about testing*. Alexandria, VA: Association for Supervision and Curriculum Development.

Schegloff, E. A. (1995). Discourse as an interactional achievement III: The omnirelevance of action. *Research on Language and Social Interaction, 28*(3), 185–211.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2): 5–8, 13, 24.

Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation, 24*, 331–345.

Shohamy, E. (2001a). *The power of tests: A critical perspective on the uses of language tests*. London: Pearson.

Shohamy, E. (2001b). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–291.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: OUP.

South Australian Curriculum and Standards Authority (2002). *South Australian Curriculum, Standards and Accountability Framework* (*SACSAF*) *English as a Second Language* (*ESL*) *Scope and Scales*. Adelaide, SA: Author.

Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D. Stevenson (Eds.), *Practice and problems in language testing* (pp. 5–21). Frankfurt: Peter Lang.

Spolsky, B. (1995). *Measured words*. Oxford: OUP.

Spolsky, B. (1997). The ethics of language gatekeeping tests: What have we learned in a hundred years? *Language Testing*, *14*(3), 242–247.

Trim, J. L. M. (1997). The proposed Common European Framework for the description of language learning, teaching and assessment. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 415–421). Jyväskylä: University of Jyväskylä and University of Tampere.

Yi'an, W. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*(1), 21–44.

L2 vocabulary acquisition theory

The role of inference, dependability and generalizability in assessment

Carol A. Chapelle
Iowa State University

Vocabulary acquisition is discussed from an assessment perspective, which provides the starting point for examining the concepts of inference, dependability, and generalizability. It argues that the construct of vocabulary needs to be defined in L2 vocabulary research if inference, dependability and generalizability are to have meaning. Inference in this context is clarified through the distinction drawn between a *construct inference* linking the theoretical and operational definitions underlying assessment, and a *theory inference* linking results from one study to the theoretical construct framework from which construct definitions are derived. Based on these concepts, I suggest that progress seems most evident in recent work on vocabulary that links acquisition, assessment, and instruction.

The title of my paper might give away my perspective. It reflects the idea that assessment could play a role in the idiosyncratic, gradual, and individual process of L2 vocabulary acquisition and therefore perhaps identifies me as a reliability-checking, truth-seeking logical positivist.[1] From this perspective, I would argue that many of the applied problems in language teaching and assessment can best be addressed by theory that is seen as true enough to be useful. I started looking at L2 vocabulary acquisition theory many years ago to develop a theory-based set of principles for evaluating constructed vocabulary responses on language tests, and have since come back to vocabulary acquisition theory for evaluating materials intended to promote incidental vocabulary learning. Both of these applied problems would be more likely to find solutions from a theory that at least specifies what L2 vocabulary knowledge[2] consists of. L2 acquisition theory should also hypothesize how such knowledge

is acquired and what constitutes more or less vocabulary knowledge, but such hypotheses need to be described within a theoretical construct framework for L2 vocabulary knowledge.

When I started looking into these issues, L2 vocabulary acquisition was characterized as an area without an accumulated body of findings from studies "concerned with establishment of a theory of the lexicon" (Gass 1988:92). Aspects of the lexicon had been studied in detail by psycholinguists (e.g., Aitchison 1987; Sternberger & MacWhinney 1988) as well as L2 researchers concerned with teaching (Nation 1990), assessment (Perkins & Linnville 1987), and acquisition (Ard & Gass 1987). However, because of the sparseness of the research and the vastly different concerns of the researchers, the work at this time did not add up to coherent perspectives that one might draw on for practice. More recently, Singleton (1999) noted that "...enough 'good' research has been published on the L2 mental lexicon in recent times to warrant a substantial review of the studies..." (p. 5). Another recent publication offers a healthy number of practice-oriented studies with implications for L2 vocabulary teaching (Coady & Huckin 1997). As lexicogrammatical patterns have recently taken a central position in much of the study of second language acquisition, it is timely to reconsider to what extent findings might ultimately add up to the type of useful theory that might guide practice.

If every study contributes individual pieces to the store of professional knowledge about vocabulary acquisition, the larger quantity of recent studies would not necessarily imply progress toward developing a theory of L2 vocabulary acquisition. Contributing to an empirically supported theory of the L2 lexicon would require researchers to interpret research results in view of the inferences underlying observed performance on research tasks as well as its dependability and generalizability – in other words, the assessment basis for the research. I will begin by explaining why I believe that vocabulary acquisition theory-for-practice rests on principles of assessment, and then, from this perspective, I explain the connection between assessment and the central concepts of this volume – inference, dependability, and generalizability. I will argue that in theory-related vocabulary assessment, inference, dependability and generalizability need to be viewed in relation to the construct definition that underlies the study of vocabulary acquisition, and the construct framework from which construct definitions are derived. Based on these concepts, I identify three directions in which progress seems evident in developing theory-for-practice in L2 vocabulary acquisition.

1. The assessment basis for vocabulary theory

Carter and McCarthy (1988) opened their book with the types of questions about vocabulary acquisition for which students and teachers would like answers. For example, "How many words provide a working vocabulary in a foreign language?" and "What are the best ways of retaining new words?" (p. 1–2). Defensible answers to these questions, like the problems I mentioned above about scoring constructed vocabulary responses and evaluating tasks for incidental vocabulary acquisition, would require a means for the researcher to determine whether or not a learner knows a given set of words.

The first question would require a researcher to identify learners who were able to perform in a criterion context, such as studying at an English-medium university, serving in a restaurant where English is used as the medium of communication, or writing reports summarizing marketing information about corn in English. Having identified the appropriate language users, the researcher would have to assess the vocabulary size by constructing and administering an *assessment* that could be demonstrated to be appropriate for estimating vocabulary size (e.g., Nation 1993). The second question would require an investigation of learners who used different strategies for learning new words. Two groups of learners might be taught different learning strategies or given different learning materials. Alternatively, a within group design would allow the researcher to gather observations of strategies that learners used for individual words and show the connection between particular strategies and word knowledge (Plass, Chun, Mayer, & Leutner 1998; Nassaji 2003). Regardless of how the learning processes for retaining new words were determined, however, the ultimate results of the research would depend on *assessment* of which words had been learned at a given point in time.

Robust answers to neither of the questions could be based on a single study, but rather evidence would be gradually supported over the course of time through multiple studies targeting the same issues and building upon one another. If one study assesses vocabulary through a multiple-choice, four-alternative test of semantic recognition knowledge, and the next study assesses vocabulary knowledge through the use of a C-test, for example, how should the results of these two studies be interpreted to address the question about successful strategies? If individual research studies are to contribute to the type of theory-for-practice that Carter and McCarthy were requesting, each needs to be interpreted in a way that appropriately informs an existing knowledge base about L2 vocabulary acquisition.

Assessment in any individual study as well as interpretation of results across studies relies on the researcher's understanding of inferential processes inherent in assessment. In the first example, appropriate interpretation of an assessment of vocabulary size requires the researcher to recognize that the results of the learners' scores on the vocabulary size test are not themselves the learners' vocabulary size. Rather vocabulary size is inferred from a summary of observed performance on the test. Similarly, in the second example, the learners' results on the vocabulary posttest are not equivalent to the learner's knowledge, but rather the learners' knowledge is inferred from results. In the third example, the results of the two tests do not address exactly the same aspects of vocabulary knowledge. Even though each test could be argued to indicate something about vocabulary knowledge, the two measure different aspects of vocabulary knowledge. In each of these cases, an interpretation of results in a way that potentially contributes to vocabulary theory requires appropriate interpretations of what the scores on the vocabulary assessments mean. In none of the examples could the vocabulary scores be interpreted to mean the level of knowledge on all of aspects of vocabulary that applied linguists would consider to be part of vocabulary knowledge. Instead, the scores obtained in research require the researcher to make inferences about vocabulary knowledge.

2. Inference in assessment

Inference in assessment refers to the logical connection that the researcher draws between observed performance and what the performance means. The process of inference has been an important part of the conceptual underpinnings of educational and psychological measurement throughout the modern history of the field (e.g., Cronbach & Meehl 1955; Messick 1989). Agreement among different schools of thought has not always existed on the kind of inferences made on the basis of scores, but researchers do agree that the score (or other summary of performance on a single occasion) is seldom in itself what the researcher is interested in. Instead, a gap exists between the score and the interpretation of interest. Recent work in assessment has helped to clarify the multifaceted inferences that are associated with the interpretation of test performance (Kane 1992; Kane, Crooks, & Cohen 1999; Kane 2001), so whereas in the past the types of inferences associated with assessment were tied to the school of thought of the researcher, today it is recognized that multiple types of inferences are associated with scores, and therefore it is up to the researcher to specify the types of inferences that are central to the interpretations that

the researcher wishes to make. In vocabulary acquisition research, at least two types – a *construct inference* and a *theory inference* – are critical for making sense of research results in a way that holds potential for informing a vocabulary acquisition theory-for-practice, and additional inferences are also relevant.

Construct inference

Virtually any researcher investigating vocabulary acquisition theory draws an inference between the observed performance summary and the construct of vocabulary that the performance is intended to signify. Figure 1 illustrates the inferential relationship between the observed performance on the vocabulary measure and the construct definition assumed as the basis for the test. For example, the pioneering studies of L2 lexicon conducted by Paul Meara in the 1970's and 1980's began with the aim of developing a more general understanding of the L2 lexicon than what could be obtained by examining error in a post hoc fashion or by probing the depth of knowledge of the idiomatic possibilities of a single lexical item. He attempted what he called a "more fundamentalist account of what interlanguage is" suggesting that researchers consider "a very crude model of what a lexicon might look like, and use this to ask some simple questions about learners' lexicons" (1984: 231). He was attempting to get away from reliance on the observed data to make explicit the object of real interest – the more general theory of the L2 lexicon.

Meara (1984) defined the construct underlying performance on a word association test as the organization of the L2 lexicon, suggesting the use of experimental data such as those obtained from word association tests to support hypotheses about lexicon organization. In a discussion of that paper pub-

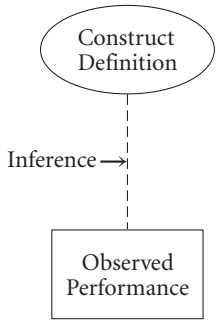

Figure 1. The inferential link between construct definition and observed performance

lished in the same volume, Michael Sharwood Smith argued that "a move to make research more fundamentalist by adopting the methods and techniques of experimental psycholinguistics will only be worthwhile if accompanied by a serious consideration of the theoretical underpinnings" and he pointed out that "experimental data cannot of themselves inform us about the nature of the learner's current mental lexicon" (1984: 239). In other words, Sharwood Smith was asking Meara to justify the inference that he was assuming between the construct of vocabulary organization and the performance on the word association tests. Methods for justification of this type of inference are the central research objective in language assessment.

The idea of the inferential link between the observed performance on a vocabulary measure and the construct of vocabulary knowledge would be readily accepted by most positivist, theory-seeking vocabulary acquisition researchers today. In other words, the fundamental idea that Meara suggested about investigating the mental lexicon is the central agenda of L2 vocabulary researchers. However, the purest positivist would insist that the construct definition would have to precede and strongly influence the selection of the test, and would question the extent to which this prescribed order was followed in studies using the word association tests for the study of lexicon organization. On this issue, I would adopt a more neo-positivist position, which would see construct definition and test development to be an iterative process, which are constrained by professional knowledge and operational realities.

Theory inference

A less acknowledged inference is the one linking the theory underlying the vocabulary construct of the test (e.g., vocabulary organization) with a broader theoretical construct framework of vocabulary knowledge. As illustrated in Figure 2, the construct that any given test is intended to measure, should typically be considered to be linked through inference to some aspects of a more complete theoretical construct framework. In research and practice, the theoretical construct framework for vocabulary knowledge is recognized to encompass multiple aspects of knowledge. For example, Nation's (2001: 27) practice-oriented perspective includes in the framework of word knowledge the receptive and productive knowledge of (1) structural word knowledge (including spoken and written forms, and word parts), (2) knowledge of word meaning (including meaning-form relations, concepts and referents, and associations), and (3) word use (including grammatical functions, collocations, and constraints on use). From an assessment perspective, a theoretical con-

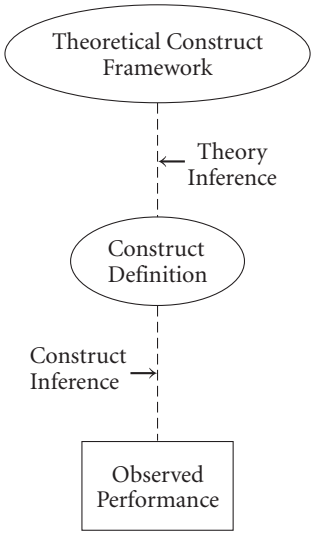

Figure 2. The inferential links between construct definition and observed performance, and between construct definition and the theoretical construct framework

struct framework would also include aspects such as vocabulary size, organization, and depth, i.e., other aspects that have been assessed as a means of detecting vocabulary acquisition (Read & Chapelle 2001).

It should be evident from these examples of dimensions of vocabulary knowledge that might comprise the construct framework that a single assessment or even multiple assessments would not succeed in assessing all dimensions of the framework within a single study or set of studies. No research has ever attempted to do so. Rather, the construct definition, which can encompass one or more aspects from the framework, offers a more realistic target for one or more assessments in vocabulary research.

The idea of a vocabulary construct framework is related to, but not the same as the framework for communicative language ability (Bachman 1990; Bachman & Palmer 1996) that has been productive in conceptualizing and interpreting language tests. The framework for communicative language ability contains "vocabulary" as a component within "grammatical knowledge," which in turn is one aspect of language knowledge that comes into play along with other factors in language use. The theoretical framework for vocabulary takes vocabulary as the starting point, detailing parameters that pertain to deployment of vocabulary knowledge, such as size, knowledge of morphological characteristics, and vocabulary strategies.

In positivist, measurement terms, constructs defined within the theoretical vocabulary framework would be within the nomothetic span (Cronbach & Meehl 1955; Embretson 1983) of constructs derived from the communicative language ability framework, but the specific relationships and their strengths would depend on the specific constructs of interest. For example, a construct such as writing ability for formal business-like letters might be defined through the relevant language knowledge (i.e., the relevant organizational and pragmatic competence) and strategic competence (e.g., assessment of audience) from a communicative language ability construct framework. A vocabulary construct such as knowledge of morphology might be hypothesized to be strongly related to the writing construct because of the need for this knowledge in deployment of precise and correct word forms in such writing. A vocabulary construct such as size might be hypothesized to be less related to the writing construct because during writing the vocabulary that the examinee needs to handle is chosen by the examinee.

The distinction between the construct definition and the theoretical framework for vocabulary has important implications in SLA research. First, if the construct definition is defined narrowly as something that can actually be assessed by one or multiple assessments, the researchers' responsibility for defining the construct of interest and justifying the inference can be taken as something that can be accomplished, or at least addressed. The idea of justifying inferences from one or multiple measures to the entire theoretical framework is overwhelming, and the assessment will fail miserably. Second, since scores on an assessment are not isomorphic with vocabulary knowledge, the theoretical relevance of any test results needs to be interpreted in view of the two inferential steps that link observed performance to a carefully delineated construct and ultimately to a theoretical framework. Third, since the theoretical framework is larger and more complex than any single construct investigated by one or multiple measures in a study, it follows that the study of all facets of the framework requires the use of a variety of types of assessments. This implication is contrary to proposals that all vocabulary researchers adopt the same methods of measurement in vocabulary acquisition research to allow for comparable results. What is needed is not a single method of measurement but defensible inferences to appropriate constructs and a single framework.

Inference, dependability, and generalizability

Recent perspectives from assessment (e.g., Kane 2001) present the ideas of inference, dependability and generalizability as closely related. According to rele-

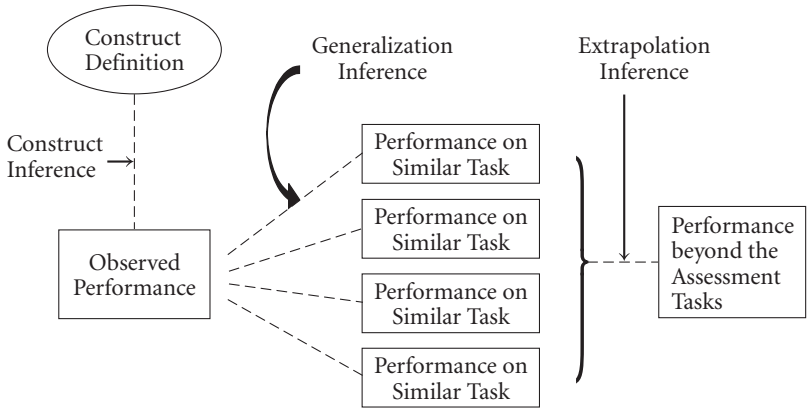

Figure 3. Construct, generalization, and extrapolation inferences embedded in the interpretation of performance on assessments

vant perspectives in assessment, the construct and theory inferences described above would be only two types of potentially many inferences associated with observed performance; an additional type of inference is the generalization that the researcher typically draws from performance on the test or research task to performance on other similar tasks. The idea is that when performance is observed on one task, it is assumed that this performance is a good representative of what would be displayed, on average, across a large number of similar tasks, as illustrated in Figure 3. In other words, the researcher is inferring similar performance across tasks. When put this way, the generalization inference encompasses the same idea as *dependability.* In other words, the inference that assumes that performance on the test task generalizes to performance on other similar tasks also assumes that performance is dependable across tasks. Generalization in applied linguistics is also used to refer to yet another inference, which more recently has been called "extrapolation." This refers to the inference linking performance on the task (or collection of similar tasks) to performance that would be displayed beyond the task(s).

Both the generalization inference (dependability) and the extrapolation inference are important for L2 acquisition theory for a number of reasons, in the interest of conciseness, however, I will focus on these inferences for the most part in terms of how they are related to the construct that the vocabulary assessment is intended to measure. In my view, the construct definition that is central to the various assessment-based inferences associated with vocabulary assessment is L2 vocabulary research.

Construct definition as central

Construct definition is central to the inferential process of assessment because of its relationship with each of the inferences required for interpretation of performance, as summarized in Table 1. Construct definition is central to the construct inference because it provides a description of the intended construct against which the test design and test performance can be evaluated. The process of justifying the construct inference is construct validation, which encompasses a range of quantitative and qualitative methods for investigating the extent to which the construct inferences are justified. Interpretation of results relative to a broader theoretical framework, i.e., the theory inference, requires identification of the specific components of the overall theoretical construct framework that the test is intended to measure. The construct definition, therefore serves as intermediary between observed performance and the theoretical construct framework, about which evidence in needed for L2 vocabulary acquisition theory. The link between the construct and the theoretical framework can be justified only so far as both are well-defined. Because of the limited recognition of this inference in L2 vocabulary research, methods for justifying its appropriateness need to be explored.

The generalization inference is connected less directly to the construct definition, but in L2 vocabulary acquisition research an important connection exists because the construct definition includes the range of knowledge that is important for specifying the range of tasks that form the relevant domain of tasks for generalization. The more narrowly and precisely the construct is defined the more likely it will be that the researcher can define and sample from the relevant universe of tasks. Dependability is another way of stating the same issue. Evidence that performance can be generalized to the appropriate

Table 1. Connections of four types of inferences to construct definition

Inference	Relationship to the construct definition
Construct	Provides a description of the intended construct against which the test design and test performance can be evaluated
Theory	Identifies the specific components of the overall theoretical construct framework that the test is intended to measure relative to the whole.
Generalization (Dependability)	Includes the range of knowledge that is important for specifying the range of tasks that form the relevant domain of tasks for generalization.
Extrapolation	Includes the type of knowledge that should clarify the domain to which the results should extrapolate.

domain suggests that data are sufficiently stable and free of error to trust them as evidence about a theoretical construct. In other words, dependability refers to having enough data and good enough data. Whether or not enough data have been yielded from performance depends on the scope of the construct definition. A construct narrowly defined as "vocabulary for reading American menus" might require fewer observations to obtain a stable sample than a more broadly defined construct such as "vocabulary ability for academic study in English-speaking North America". Similarly, whether the data are good enough depends in large part on what the data are supposed to reflect. Bachman (1990) put this concept very clearly: "The concerns of reliability and validity can . . .be seen as leading to two complementary objectives. . .: (1) to minimize the effects of measurement error, and (2) to maximize the effects of the language abilities that we want to measure" (1990: 161). In later work Bachman and Palmer (1996) refer to "measurement error" as "unmotivated variation." This defines the issue even more clearly: in order to distinguish error or unmotivated results from construct relevant or motivated ones, a clear construct definition is essential.

The extrapolation inference is related to construct definition because the way the construct is defined should help to clarify the domain to which the results should extrapolate. Again, the issue is the scope of the construct definition. For example, a construct stated as "vocabulary organization" does not imply any particular domain of extrapolation. Instead, if the construct definition refers to vocabulary organization as it is structured during particular activities (Votaw 1992) the general specification of the activities offers guidance to specify the appropriate domain of extrapolation. Another example would be the contrast between the construct of vocabulary size in general as I have referred to it above and the more specific receptive vocabulary size for university studies in Dutch universities (Hazenberg & Hulstijn 1996).

3. Progress for vocabulary acquisition research

Having described the problem of developing L2 vocabulary acquisition theory in terms of the inferences that researchers need to be able to make from observed performance data, I can now assess the extent to which the more plentiful recent work in this area can be seen as representing progress. Three areas of progress are apparent.

Probing construct definition

In view of the central role of construct definition for all inferences, the explicit discussion of vocabulary knowledge as a construct seems to mark progress. Meara's suggestion to look at a "crude model of what a lexicon might look like" called for considering vocabulary as a construct. Telchrow (1982) distinguished two constructs – receptive and productive vocabulary ability – a distinction that has continued to be useful for in the work of Laufer (1998), which looks at the development of these different but related constructs. Other research attempted to expand on the construct that had been defined primarily through the dimensions of size and mental organization, and with lists of characterizing word knowledge. Ard and Gass (1987), for example, explored the syntactic dimensions of the construct, and Read (1993, 1998) attempted to develop a measurable construct of vocabulary depth. Chapelle (1998) made explicit three perspectives from which a construct definition for vocabulary can be developed: A trait definition is specified as general knowledge intended to be relevant across many domains, a behaviorist definition is stated in terms of the contexts in which performance is expected to be evident, and an interactionalist definition includes the relevant knowledge delimited by the contexts of interest. Rather than expanding on or detailing a composite vocabulary construct, the three perspectives articulate different perspectives that can be taken to in defining vocabulary as a construct.

In today's work, researchers have not gone so far as to describe the construct they are investigating in terms of the three approaches, but nevertheless, it seems that we have passed the time when vocabulary ability was a fuzzy concept for which some evidence could be gained from any learner performance that happened to be on hand. Instead, researchers are investigating specifically-defined aspects of the construct such as productive knowledge of derivative morphology (Schmitt & Zimmerman 2002) with tests that are argued to support valid construct inferences. In like fashion some of the vocabulary tests that have been developed target a specific construct such as depth of vocabulary knowledge (Wesche & Paribakht 1996). These steps forward might eventually benefit L2 vocabulary theory through increased clarity in specifying the construct definition underlying a specific test as distinct from the theoretical construct framework.

Validation studies

Perhaps the most obvious sign of progress in L2 vocabulary research is the increase in validation studies of measures that are used in both L2 teaching and research. A validation study focuses specifically on producing evidence for the appropriateness of the inferences made from assessment scores. Two such validation studies were reported recently in *Language Testing* concerning vocabulary measures that have been in use for many years in SLA research. One study found reason to question the inferences associated with the Yes/No vocabulary test (Beeckmans, Eychmans, Janssens, Dufranne, & Van de Velde 2001), and another presents a range of validity evidence concerning the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham 2001). These two examples may mark the beginning of a positive trend toward empirical validation of the vocabulary tests used in L2 studies – a process which is not altogether new (e.g., Cronbach 1942, 1943; Feifel & Lorge 1950), but which has not been undertaken for measures used in L2 vocabulary studies, which have tended to rely on researchers' judgments of what tests might measure (e.g., Kruse, Pankhurst, & Sharwoor Smith 1987; Laufer 1990; Palmberg 1987; Singleton & Little 1991). Although researchers' judgments are not irrelevant, they should not constitute the entire validity argument.

Validation studies hopefully mark a beginning for the accumulation of validity evidence that can be developed into more thorough validity arguments along the lines suggested by researchers in educational measurement (Messick 1989; Kane 2001). Such an argument needs to accumulate evidence about a test or test method from multiple sources to provide evidence about the construct underlying test performance, as illustrated by Chapelle (1994). Such an argument allows the researcher to draw on the relevant theoretical and empirical rationales from linguistics (e.g., Anshen & Aronoff 1988; Wray 2002) and cognitive psychology (e.g., Schwanenflugel & Shoben 1985; Tyler & Wessels 1983) to develop a construct definition with appropriate depth and to weigh the justification of links between the construct definition and empirical results. Developing a validity argument for L2 vocabulary in this way forces the researcher to state an explicit L2 vocabulary construct definition.

Salience of L2 vocabulary assessment

One of the most encouraging signs of progress in the work on vocabulary acquisition is the link being made among those working in teaching, research and assessment, as demonstrated in two recent books. In Paul Nation's (2001)

Learning Vocabulary in Another Language, L2 vocabulary acquisition is surveyed, particularly as it pertains to language teaching. The book contains tests that have been developed and used in the classroom and for research *and* it makes reference to the validation research on the tests (e.g., Schmitt, Schmitt, & Clapham 2001). The book reviews some issues in vocabulary assessment and explicitly discusses limitations of vocabulary tests with statements such as the following: "Direct tests of vocabulary size . . . do not show whether learners are able to make use of the vocabulary they know; and they do not measure learners' control of essential language learning strategies like guessing from context, dictionary use and direct vocabulary learning" (Nation 2001:382).

Read's (2000) book, *Assessing Vocabulary*, summarizes the state of the field with respect to construct definition while showing the need for a clear construct definition to guide development and evaluation of vocabulary tests. Read's approach to vocabulary assessment takes a step toward helping vocabulary researchers understand the links between test design and construct definition. He identifies characteristics of vocabulary measures that help to define any given measure in terms that can be used for comparison. He uses a framework including three dimensions (i.e., discrete/embedded, selective/comprehensive, and context-independent/context-dependent), and then uses these dimensions to describe a variety of vocabulary measures. Vocabulary assessments can therefore be discussed within a common framework. Having a means of distinguishing among different types of measures is an essential first step toward avoiding the confusion ensuing from attempts to compare results from two different L2 vocabulary tests measuring different constructs.

The next step in the process of understanding inferences, generalizeability and dependability is to relate Read's assessment framework to the perspectives of construct definition. For example, the discrete/selective/context independent vocabulary tests that have been used for so many years tend to be appropriate for making inferences about a trait-type construct definition. As vocabulary assessment moves closer to the ideals of other areas of language assessment with embedded, comprehensive, and/or context dependent measures, the construct about which inferences are made moves away from the trait-type to a more interactionalist-type. Read and Chapelle (2001) attempt a preliminary analysis of some vocabulary assessments along these lines. It remains to be seen if this approach is useful for vocabulary researchers attempting to address construct and theory inferences in L2 vocabulary research. But in the meantime, combining perspectives from acquisition, teaching and assessment, as they pertain to the issues of construct definition and validation, Read is able to demonstrate, as Schoonen puts it in his review of Read's book in *Language*

Testing, "that there is much more to vocabulary assessment than the traditional multiple choice test" (Schoonen 2001:118).

4. Conclusion

The landscape in L2 vocabulary research is radically different today than it was 15 years ago. The three areas of progress I mentioned center on inferences that depend on the construct definition underlying a test, and in particular the construct inference has been the focus of much attention. In contrast, much less progress is apparent in studying the theory inference. Huckin and Haynes (1993), for example, suggest that "researchers would do well. . . to try to assemble converging evidence from multiple methods. Single methods tend to yield one-dimensional perspectives, which are simply inadequate for a subject as complex as second language reading and vocabulary learning" (p. 297). In addition to this good measurement advice, advice in needed about how data from multiple methods should be interpreted. Do multiple methods of measurement speak to the construct inference, to the theory inference, or to both? The idea that (at least) two inferences are needed to link observed performance data to a theoretical construct framework has not been recognized explicitly in the L2 vocabulary research literature. Such recognition rests on a clear specification of the inferences that the researcher makes to explain observed performance.

Notes

1. Assessment does not have to be conducted from a logical positivist's perspective; however, most readers would associate assessment with logical positivism. Moreover, the conceptual apparatus of inference, dependability and generalizability was developed within this perspective. See Lynch (2003) for discussion of logical positivist vs. interpretivist perspectives on assessment.

2. I am using the term "knowledge" throughout because this is the term that is typically used in L2 vocabulary research. In fact, in most applied settings, what would be of interest might more arguably be called "ability" which includes both knowledge and the processes for putting the knowledge to use, by analogy with "communicative language ability" as defined by Bachman (1990).

References

Aitchison, J. (1987). *Words in the mind: An introduction to the mental lexicon*. New York: Basil Blackwell.

Anshen, F. & Aronoff, M. (1988). Producing morphologically complex words. *Linguistics, 26,* 641–655.

Ard, J. & Gass, S. (1987). Lexical constraints on syntactic acquisition. *Studies in Second Language Acquisition, 9,* 233–351.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: OUP.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice.* Oxford: OUP.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the Yes/No vocabulary test: Some methodological issues in theory and practice. *Language Testing, 18*(3), 235–274.

Carter, R. & McCarthy, M. (1988). *Vocabulary and language teaching.* London: Longman.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Second language acquisition and language testing interfaces* (pp. 32–70). Cambridge: CUP.

Chapelle, C. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10*(2), 157–187.

Coady, J. & Huckin, T. (Eds.). (1997). *Second language vocabulary acquisition*. Cambridge: CUP.

Cronbach, L. J. (1942). Analysis of techniques for diagnostic vocabulary testing. *Journal of Educational Research, 36,* 206–217.

Cronbach, L. J. (1943). Measuring knowledge of precise word meaning. *Journal of Educational Research, 36,* 528–534.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Embretson, S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197.

Feifel, H. & Lorge, I. (1950). Qualitative differences in the vocabulary responses of children. *Journal of Educational Psychology, 41,* 1–18.

Gass, S. M. (1988). L2 Vocabulary Acquisition. *Annual Review of Applied Linguistics, 9,* 92–106.

Hazenberg, S. & Hulstijn, J. H. (1996). Defining a minimal receptive second language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics 17*(2), 145–163.

Huckin, T. & Haynes, M. (1993). Summary and future directions. In T. Huckin, M. Haynes, & J. Coady (Eds.), *Second language reading and vocabulary learning* (pp. 289–298). Norwood, NJ: Ablex.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

Kruse, H., Pankhurst, J., & Sharwood Smith, M. (1987). A multiple word association probe in second language acquisition research. *Studies in Second Language Acquisition, 9,* 141–154.

Laufer, B. (1998). The development of passive and active vocabulary in a second language. *Applied Linguistics, 19*(2), 255–271.

Laufer, B. (1990). "Sequence" and "Order" in the development of L2 lexis: Some evidence from lexical confusions. *Applied Linguistics, 11*(3), 281–296.

Lynch, B. (2003). *Language assessment and program evaluation.* New York: Columbia University Press.

Meara, P. (1984). The study of lexis in interlanguage. In A. Davies, C. Criper, & A. P. R. Howatt (Eds.), *Interlanguage* (pp. 225–240). Edinburgh: Edinburgh University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). NY: Macmillan.

Nassaji, H. (2003). L2 vocabulary learning from context: strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *TESOL Quarterly*, *37*(4), 645–670.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: CUP.

Nation, I. S. P. (1993). Using dictionaries to estimate vocabulary size: Essential but rarely followed procedures. *Language Testing, 10,* 27–40.

Nation, I. S. P. (1990). *Teaching and learning vocabulary in another language.* New York, NY: Heinle and Heinle.

Palmberg, R. (1987). Patterns of vocabulary development in foreign language learners. *Studies in Second Language Acquisition, 9,* 201–220.

Perkins, K. & Linnville, S. (1987). A construct definition study of a standardized ESL vocabulary test. *Language Testing, 4*(2), 126–141.

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology, 90*(1), 25–36.

Read, J. (2000). *Assessing vocabulary*. Cambridge: CUP.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Lawrence Erlbaum.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10,* 355–371.

Read, J. & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*, 1–32.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55–88.

Schmitt, N. & Zimmerman, C. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*(2), 145–171.

Schoonen, R. (2001). [Book Review.] *Language Testing, 18*(1), 118–125.

Schwanenflugel, P. L. & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language,* 24, 232–252.

Sharwood Smith, M. (1984). Discussion. In A. Davies, C. Criper, & A. P. R. Howatt (Eds.), *Interlanguage* (pp. 236–239). Edinburgh: Edinburgh University Press.

Singleton, D. (1999). *Exploring the mental lexicon*. Cambridge: CUP.
Singleton, D. & Little, D. (1991). The second language lexicon: Some evidence from university-level learners of French and German. *Second Language Research, 7,* 62–81.
Sternberger, J. P. & MacWhinney, B. (1988). Are inflected forms stored in the lexicon? In M. Hammond & M. Noonan (Eds.), *Theoretical morphology: Approaches in modern linguistics* (pp. 101–116). San Diego, CA: Academic Press.
Telchrow, J. M. (1982). A survey of receptive versus productive vocabulary. *Interlanguage Studies Bulletin, 6,* 3–33.
Tyler, L. K. & Wessels, J. (1983). Quantifying contextual contributions to word recognition processes. *Perception and Psychophysics, 34,* 409–420.
Votaw, M. C. (1992). A functional view of bilingual lexicosemantic organization. In R. J. Harris (Ed.), *Cognitive processing in bilinguals,* (pp. 299–321). New York, NY: Elsevier Science Publishers.
Wesche, M. & Parabut, T. S. (1996). Assessing second language vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review, 53,* 13–39.
Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: CUP.

Beyond generalizability

Contextualization, complexity, and credibility in applied linguistics research

Patricia A. Duff
University of British Columbia

This chapter discusses the issues of inference and generalizability in qualitative applied linguistic research primarily. The following themes are also explored in relation to generalizability (or transferability) in case studies, ethnographic classroom research, and other forms of qualitative research: (1) seeking *contextualization*; (2) understanding the *complexity* of behaviors, systems, and beliefs through triangulation and thick description; and (3) establishing the *credibility* of interpretations drawn from research and their importance for applied linguistic theorizing more generally. I conclude that both quantitative and qualitative applied linguistic research should seek to maximize the awareness and importance of contextualization, complexity, and credibility in studies as well as analytic or naturalistic generalizability, as we engage in rigorous, systematic, and meaningful forms of inquiry and interpretation.

Introduction

Research in education and the social sciences has long been concerned with the basis for inferences and conclusions drawn from empirical studies, and applied linguistics is no exception. With the emergence of qualitative, mixed-method, and innovative new approaches to research in recent decades, issues connected with validity in research and testing have received renewed attention by applied linguists, not only about how to design and interpret one's own studies in order to make legitimate claims, but also how to interpret others' research or the tools and products of research, such as tests, typologies, and so on. One particular area of concern is the nature and scope of insights that can be generated from qualitative research such as case study, ethnography, narrative inquiry,

conversation analysis, and so on, as well as within more familiar quantitative research paradigms or inquiry traditions, especially when small sample sizes are involved. Seeking and reaching consensus regarding criteria for conducting and evaluating high quality research within each tradition has therefore become a priority of late (e.g., Chapelle & Duff 2003; Edge & Richards 1998; Lazaraton 2003).

This chapter discusses inference and generalizability as they apply to *qualitative* applied linguistic research primarily, while conceding that the terms qualitative and quantitative are overstated binaries when describing contemporary so-called "new-paradigm" research designs. First, I define inference and generalizability as they are often understood and used in quantitative research and consider their relevance to qualitative research. I then explore such themes as *contextualization, complexity,* and *credibility* in relation to generalizability in most qualitative research as fundamental ways of broaching validity. Case studies and ethnographies in second language acquisition (SLA) and second language education (SLE) are selected to illustrate these principles. I conclude that both quantitative and qualitative applied linguistic research should seek to maximize the awareness and importance of contextualization, complexity, and credibility as well as analytic or naturalistic generalizability (however these concepts are taken up) as we promote rigorous, systematic, and meaningful forms of inquiry and as we consider the inferences and interpretations that can be drawn from our research.

Inference

The aim of research is to generate new insights and knowledge – in other words, to make various kinds of inferences based on observations. Quantitative research often stresses the importance of inference. In that paradigm, inference may refer to the degree to which we infer from people's observable behaviors aspects of their underlying competence or knowledge systems, or it may refer to the nature of claims that are inferred from various kinds of evidence (e.g., about the relationships among variables). Inference, unlike generalizability, is a term that is seldom addressed explicitly or in a technical sense in meta-methodological discussions of qualitative research, as a perusal of the subject indexes of many qualitative research methods textbooks reveals (e.g., Crabtree & Miller 1999; Creswell 1998; Denzin & Lincoln 2000; Eisner & Peshkin 1990; Holliday 2002; Merriam 1998; Miles & Huberman 1994; Neuman 1994; Silverman 2000). Even in quantitative or mixed-methods

textbooks, the term is mainly used in conjunction with particular types of statistics.[1] Inferential statistics, such as t-tests or analysis of variance, make it possible to draw certain kinds of conclusions about data in relation to research questions (e.g., causality), and especially about the relationship between the sample data and the characteristics of the larger population (Brown & Rogers 2002; Fraenkel & Wallen 1996; Gall, Borg, & Gall 1996; Palys 1997). Inference, therefore, is closely related to the notion of generalizability: namely, whether results are generalizable to a larger group or to theoretical principles, that is, whether we can infer generality. In some respects then, inference and generalization can refer to the same process, since generalization is a kind of inference. However, inference refers to a broader cognitive process connected with logical reasoning, generality being just one sort of reasoning. Thus, although inference is a concept not typically discussed at length – or even in passing – in most qualitative research, it is a form of reasoning that is implicit in the presentation and interpretation of research within any paradigm. For example, qualitative research in language education seeks to draw inferences about such topics as the values, conditions, linguistic and sociocultural knowledge, and personal experience that underpin observable behaviors in order to understand how and why people behave or interact in certain ways, how they interpret their behaviors and situations, how learning proceeds, what instructional processes are deemed effective, the characteristics of learning cultures, the attributes of certain kinds of learners and teachers, and so on. Instead of inference, most qualitative methodology textbooks discuss processes of *interpretation*, a kind of inference, a search for patterns and understandings, which is central to the meaning-making in qualitative research.

Generalizability in quantitative vs. qualitative research: Problematics and possibilities

Generalizability, a crucial concept in positivist (generally quantitative) experimental research, aims to establish the relevance, significance, and external validity of findings for situations or people beyond the immediate research project. That is, it is part of the process of establishing the nature of inferences that can be made about the findings and their applicability to the larger population and to different environmental conditions and to theory more universally. Generalizability (or generality as some describe it, e.g., Krathwohl 1993), while typically discussed in connection with inferences about populations (whether the observations about sampled individuals can be generalized

to others in the same population), can also involve the ability to generalize (effects) to "treatments, measures, study designs, and procedures other than those used in a given study," according to Krathwohl (p. 735). Sampling procedures are one of several elements (e.g., research design) that affect the kinds of inferences that can be drawn about generality. It is commonly accepted that quantitative research, with appropriate sampling (random selection, large numbers, etc.), research design (e.g., counterbalancing of treatments, ideally with a control group, pre-post measures, and careful testing and coding procedures), and inferential statistics where appropriate, has the potential to yield generalizable results.

However, carefully controlled research may nevertheless provide inadequate contextualization of the study, the participants, the tasks/treatments, and so on, and therefore may be less easily generalized than might otherwise be the case. As Gall et al. (1996) point out, "generalizing research findings from [an] experimentally accessible population to a target population is risky" (p. 474). The two populations must be compared for crucial similarities. In addition, "if the treatment effects can be obtained only under a limited set of conditions or only by the original researcher, the experimental findings are said to have low ecological validity" (p. 45), and thus low external validity as well. Safeguards must be in place both in designing and carrying out the experiments and also in reporting the results. Gall et al. (1996) provide a list of factors associated with external validity in experiments that researchers need to take into account (see Table 1), listed under the headings of population validity and ecological validity (after Bracht & Glass 1968).

Key sociocultural variables such as institutional context, first language (L1) background or the relationship between first and second language/culture, and other characteristics of the sample/population, the tasks, and even the relationship between the researcher and those researched might be underspecified or omitted, because of space limitations or because the variables are not considered important or central to the study. The generalizability of findings to wider populations and contexts can inadvertently be reduced as a result – regardless of the claims made by the researchers about generality. In a field as international, interlingual, and intercultural as SLA or SLE, the sociocultural, educational, and linguistic contexts of research are of great importance (both macro-level social contexts and micro-level discursive/task contexts, Duff 1995). To date, most of the published research conducted in TESOL, for example, has taken place in American university or college programs with students within a particular range of proficiency and educational preparedness that may or may not be easily generalized to other types of programs (e.g., with children) or

Table 1. Generalizability and quantitative vs. qualitative research

	Quantitative	Qualitative
Research designs/methods	– quasi-/experiments, correlations, surveys, regressions, factor analyses, etc.	– case study, ethnography, conversation analysis (CA), other micro-discourse analyses, interview research, document analysis, or some combination of these
View of generalizability (external validity)	– emphasis on external validity (e.g. in experiments; Gall et al. 1996: 466): 1. population validity 2. ecological validity, with attention to: – explicit description of experimental treatment – "representative design" (reflects real-life environments and natural characteristics of learners) – multiple-treatment interference – Hawthorne effect – novelty and disruption effects – experimenter effect – pretest sensitization – posttest sensitization – interaction of history and treatment effects – measurement of dependent variable – interaction of time of measurement and treatment effects Krathwohl (1993): explanation generality, translation generality, demonstrated generality, restrictive explanations (conditions) eliminated, replicable result	– limited relevance, in traditional views, to the goals of qualitative research; validity is described instead in terms of "transferability," catalytic and ecological validity, credibility, dependability, confirmability, etc.; "validity" and "reliability" are often not compartmentalized (Gall et al. 1996) – some (positivist) qualitative researchers believe that through the aggregation of qualitative studies from multiple sites (e.g. case studies through a case survey method, through "meta-ethnography" or cross-case "translation," and the comparative method) generalizations may be warranted; others see generalizability in terms of "fit" between one study/situation and another, based on thick description (Schofield 1990)

Table 1. (*continued*)

	Quantitative	Qualitative
Types of generalization	– to populations, settings, conditions, treatments – to theories/models (hypotheses)	– to populations (similar cases/contexts; also known as case-to-case translation), Lincoln & Guba (1985), based on similarities across situations/contexts, and enhanced by the representativeness of sites/cases/situations studied – to theories/models (this is the more common application; also referred to as "analytic generalizability," Firestone 1993)
Strengths	– multiple occurrences of phenomenon/effects in relatively controlled setting; quantification helps establish clear trends and relationships; statistical inferences or extrapolation of results to defined wider population possible (especially with random or probability sampling) – normally based on data from large sample populations – internal, external validity and other types of validity often demonstrated convincingly; attention paid to reliability of coding, testing, etc.	– great potential for rich contextualization, accounting for complexity of social/linguistic phenomena – potential to resonate with readers because of accessibility, description, narrative genre and examples – in critical/feminist research, the potential to move readers to action to create change; i.e. "catalytic validity" (Lather 1991) – more opportunities to document ecological or internal validity than in much quantitative research; but less potential or desire to make claims of external validity – based on theoretical or purposive sampling – transferability of findings is determined not only by researcher but also by reader and by the typicality, representativeness or depth of description of case/situation – focus on multiple meanings and interpretations
Way of strengthening further	– using ecologically valid tasks, settings, procedures; addressing underlying constraints on generalizability – documenting contexts, sampling, procedures, instruments, etc. in as much detail as possible	– having an aggregation of multiple-case or multiple-site studies; triangulation – thick description – include participants' own judgment of generalizability/representativeness or typicality (Hammersley 1992); corroboration from other studies (cf. Maxwell 1996), in addition to observations

Table 1. (*continued*)

	Quantitative	Qualitative
Potential problems	– over-emphasis on external validity in some cases, at the expense of internal validity (e.g., task-based research, with strangers, or using artificial languages, contrived tasks): How far can findings about interaction patterns be generalized to "normal/natural" learning conditions? – failure to replicate or follow up on studies with different populations and in different contexts may lead to de facto generalization – too often, the research focuses on only the analysts' analysis/perspectives and not those of participants or other observers (although coding is usually strengthened by getting 2nd raters, using operational definitions/constructs, etc.) – statistical inferences may be based on inappropriate statistical procedures for types of data involved (cf., Hatch & Lazaraton 1991)	– over-emphasis on possibly *atypical*, *critical*, *extreme*, *ideal*, *unique*, or *pathological* cases, rather than typical or representative cases (e.g., Genie, Alberto, Wes; i.e. in terms of fossilization, critical period studies, exceptionally good or ineffective language learners) – "telling cases" may not always be highly representative cases; they may be very helpful in providing insights into SLA but conditions or insights may not apply broadly to others; e.g., autobiographical cases of metalinguistically sophisticated learners (Schmidt & Frota 1996) who may be atypical – tendency to generalize widely to theory (acculturation model, notice-the-gap principle), despite disclaimers, with few similar case studies conducted (i.e. in lieu of replication, multiple-case-studies) – observations may lack relevance to field more widely, outside of immediate context – need for enduring key themes, constructions, or relationships among factors that will be helpful to other researchers in other contexts; going beyond particularistic local observations to more general trends – rigorous evidence or authentication may seem to be lacking (Edge & Richards 1998) but goal should be to "interpret and . . . offer a [context-specific] understanding" (p. 350) – see Lazaraton's (2003) discussion of "criteriology" and qualitative social research

in countries with very different educational systems, cultures, histories, and economies (e.g., in EFL regions).

The more controlled and laboratory-like the SLA studies (e.g., Hulstijn & DeKeyser 1997), often using very contrived tasks (even involving artificial languages or invented images in some cases in order to control for prior knowledge), the less generalizable the findings are, in my view, either to larger populations from which samples are drawn or to broader understandings of language teaching, learning, or use in classrooms or other naturally occurring settings. That is, it may be difficult and unwise to generalize from behaviors of unfamiliar pairs of interlocutors doing unclassroom-like research tasks under laboratory-like conditions to how language learners, as familiar classmates with their own history of interacting with one another and undertaking tasks, would do classroom tasks or engage with interlocutors in natural, non-experimental settings.[2] Furthermore, while the research may speak to issues of how they would engage in one particular type of task, it does not shed light on how curriculum can be developed linking such tasks in meaningful, educationally sound ways. Most SLA studies continue to examine L2 learners of English or other European languages and much less often European-L1 learners of non-European L2s (e.g., English learners of Chinese or Arabic; see Duff & Li 2004, on this point). They have also privileged a small set of fairly basic tasks associated with communicative language teaching (e.g., spatial "spot-the-difference" or "plant the garden" tasks), and have not explored other instructional contexts such as EFL instruction to the same extent, and far too seldom investigate interactions or language development over an extended period of time. These issues, in my view, also reduce the generalizability and utility of the findings of such studies, no matter how rigorously they are conducted. The onus is on researchers in quantitative studies to convincingly demonstrate the external validity of their findings (if that is their objective), rather than take it for granted that generalizability is possible in quantitative research but categorically impossible in qualitative research.

Table 1 captures some basic differences between quantitative vs. qualitative research, especially in relation to generalizability and the strengths and weaknesses inherent in each paradigm with respect to validity concerns, more broadly. Here I summarize some of the most commonly cited differences. Whereas quantitative research emphasizes both internal and external validity (or generalizability), in addition to reliability, qualitative research, especially postpositivist, naturalistic, interpretive studies, typically emphasize elements associated with a combination of *internal validity* and *reliability* (to borrow terms from quantitative research). Internal validity in quantitative research,

like its counterpart in qualitative research, is related to the credibility of results and interpretations, as a function of the conceptual foundations and the evidence that is provided. As Krathwohl (1993) puts it, internal validity in quantitative research is related to the study's "conceptual evidence linking linking the variables," supported by "empirical evidence linking the variables" – demonstrated results, the elimination of alternative explanations, and judgments of the overall credibility of the results (p. 271). In qualitative research, internal validity is addressed by means of contextualization; thick description; holistic, inductive analysis; triangulation (or "crystallization," to use a more multifaceted metaphor; Richardson 1994); prolonged engagement; ecological validity of tasks; and a recognition of the complex and dynamic interactions that may exist among factors; as well as the need for the credibility or trustworthiness of observations and interpretations (Davis 1995; Watson-Gegeo 1988). Thick description, one of the most touted strengths of case study and ethnography (but not of conversation analysis or certain other kinds of qualitative inquiry), may draw on the following sources of information: documentation, archival records, interviews, direct observations, participant-observation, and physical artifacts (Yin 2003). Gall, Gall, and Borg (2003) suggest that a suitably thick description of research participants and concepts allows "readers of a case study report [to] determine the generalizability of findings to their particular situation or to other situations" (p. 466). The aim is to understand and accurately represent people's experiences and the meanings they have constructed, whether as learners, immigrants, teachers, administrators, or members of a particular culture.

Qualitative research in education, according to Schofield (1990), first began to address issues of generalizability because of large-scale, primarily quantitative, multi-method program evaluation research in the 1980s and 1990s that incorporated significant qualitative components and yet, given the overarching quantitative structure, still framed discussions in terms of generalizability. She observed that a general "rapprochement" of the two major paradigms since then, emphasizing their complementarity as opposed to fundamental incompatibility, has further prompted researchers to examine reliability and validity or their proxies. Although it is often said that qualitative research is neither interested in nor able to achieve generalizability or to generate causal models or explanations, there are in fact many diverging opinions on this issue, as we will see in what follows (Bogdan & Biklen 1992).

In applied linguistics, in general, qualitative research seeks to produce an in-depth exploration of one or more sociocultural, educational, or linguistic phenomena and, in some cases, of participants' and researchers' own position-

ality and perceptions with respect to the phenomena (Davis 1995). Rather than understand a phenomenon in terms of its component parts, the goal is to understand the whole as the sum or interaction of the parts (Merriam 1998). Generalizability to larger populations, in the traditional positivist sense, or prediction is not the goal. As Stake (2000) observes with respect to case studies, "the search for particularity [in a case, or a biography] competes with the search for generalizability" (p. 439).

Indeed, for many qualitative researchers, the term generalizability itself is considered a throw-back to another era, paradigm, ethos, and set of terminology (or discourse) in research. Schofield (1990) stated it as follows: "[t]he major factor contributing to the disregard of the issue of generalizability in the qualitative methodological literature appears to be a widely shared view that it is unimportant, unachievable, or both" (p. 202). She attributes this disinterest in part to the origins of much (ethnographic) qualitative research to cultural anthropology, which studies other cultures for their intrinsic value: for revealing the multiple but highly localized ways in which humans live. Cronbach (1975, cited in Merriam 1998) suggested that social science – and not just qualitative – research should not seek generalizability anyway: "When we give proper weight to local conditions, any generalization is a working hypothesis, not a conclusion" (Merriam 1998:209). Note that the association of the replacement term "working hypotheses" for "generalization" by some qualitative researchers is not universally accepted though because it downplays the affective, interactive, emergent nature of perspective-sharing in qualitative research and again uses terminology associated with quantitative research (Merriam 1998).

Instead, the assumption is that a thorough exploration of the phenomenon/a in one or more carefully described contexts – of naturalistic or instructed L2 learners with various attributes, classrooms implementing a new educational approach, or diverse learners integrated within one learning community – will be of interest to others who may conduct research of a similar nature elsewhere. Other readers may simply seek the vicarious experience and insights gleaned from gaining access to individuals and sites they might otherwise not have access to. Stake (2000 and elsewhere) refers to the learning and enrichment that proceeds in this way as "naturalistic generalization" – learning from others' experiences. Adler and Adler (1994) suggest that researchers' writing or reporting style itself contributes a great deal to the impact of qualitative research on readers and their perceptions of its credibility and authenticity and that researchers should strive to achieve what they call "verisimilitude" in their

writing – "a style of writing that draws the reader so closely into the subjects' worlds that these can be palpably felt" (p. 381).

Another term commonly replacing generalizability in the qualitative literature is *transferability* – of hypotheses, principles or findings (Lincoln & Guba 1985). Transferability (also called comparability) assigns the responsibility to readers to determine whether there is a congruence, fit, or connection between one study context, in all its complexity, and their own context, rather than have the original researchers make that assumption for them. Still, some qualitative researchers find the concept of transferability to be too similar in focus to generalizability to be a useful departure from traditional views and makes too much of the need for similarity or congruence of studies (e.g., Donmoyer 1990). They feel that difference, in addition to similarity, helps sharpen and enrich people's understandings of how general principles operate within a field beyond what the notion of transferability suggests. Also, rather than seeking "the correct interpretation," they would aim to broaden the repertoire of possible interpretations and narratives of human experience. Qualitative research, in this view, provides access to rich data about others' experience that can facilitate understandings of one's own as well as others' contexts and lives, both through similarities and differences across settings or cases.

In summary, in both quantitative and qualitative research, it is important to establish the basis for one's findings and interpretations, but in the latter, strictly causal inferences, statistical inferences or universal "laws" are not normally sought or warranted. Also, the nature of ontological "truth" claims sought or made are different in that in quantitative research it is assumed that there is an external truth or reality to be discovered whereas most qualitative research (even positivist case study methodology as advocated by Yin 2003) would assume that there are multiple possible "truths" to be uncovered or (co-)constructed, which may not always converge. However, qualitative research is not only intended to be of limited, highly local significance without addressing issues of broader relevance and meaning. Much qualitative research does seek to provide generalizations at an abstract conceptual or theoretical level, what Firestone (1993) refers to as *analytic generalization*, well beyond the details of cases (e.g., related to models of SLA or SLE). Some qualitative research, furthermore, seeks to provide causal explanations about observations (Miles & Huberman 1994), in addition to thick description. Thus, generalizability and the nature of explanations that research can support have been problematized in both qualitative and quantitative research in our field.

Inference and generalizability in qualitative research: Case study and ethnography

In case studies and classroom ethnographies, it is necessary for researchers to acknowledge the delimited or context- (or culture-) bounded nature of their observations and findings while at the same time providing rich, detailed descriptions of the sites, participants and other objects of inquiry. Ordinarily, a small number of main themes or different participant profiles emerging from iterative data analysis are discussed, along with complex interactions among variables or factors linked with observed behaviors or situations. The observations are typically contextualized both within socio-educational settings and within the larger theoretical research literature and the set of issues that motivated the study. If done well, the metaphors, models, themes, and constructs that emerge may influence and inform subsequent research.

Most case studies and ethnographies are not characterized or evaluated in terms of their generalizability, but rather in terms of more social, ethical, textual, and data-analytic concepts such as: *contextualization* or *contextual completeness, thick description*; *prolonged engagement* (longterm observation), *complexity, triangulation, multivocality* (multiple voices or perspectives), *credibility, relevance, plausibility*, (*researcher*) *positionality and reflexivity, trustworthiness, authenticity, usefulness*, and *chains of evidence*. Most of these terms relate to practices of providing convincing evidence and arguments, from a variety of sources, to support interpretations (or inferences). Altheide and Johnson (1994) call these aspects of "interpretive validity." I discuss these concepts in the context of some recent SLA/SLE research below and review ways in which some researchers recommend that generalizability or transferability be enhanced in qualitative inquiry.

Contextualization, complexity, and credibility

Contextualization

A crucial aspect of qualitative research in applied linguistics is understanding and documenting the research context (Johnson 1992): that includes the larger sociopolitical or historical context (where relevant), the participants and their interests, the tasks or instructional practices used and participants' understandings or views of these, in some cases, and how the research itself, whether inside a classroom or in a research office of some kind, creates a special sociolinguistic context, system, or ecology that is temporally as well as socially and discursively situated (van Lier 1988, 1997). Understanding learning communi-

ties as ecologies or organic systems is increasingly being stressed now (Kramsch 2002), as is the importance of local contexts, including discourse contexts, and cultures of language learning and use (Breen 1985; Duranti & Goodwin 1992; Holliday 1994). Even in non-ethnographic inquiry, task-based research can be analyzed with a view to understanding the kinds of contexts that are being created by tasks and the impact of those contexts on the generation and interpretation of findings over time (Coughlan & Duff 1994; Duff 1993b).

In my studies examining educational systems in transition as a result of changing social/linguistic demographics, L2 policies, curricula, and so on, documenting these larger contexts in which more focused observations are situated and understanding macro-micro interfaces to a greater extent has been key (Duff 1995; Duff & Early 1996). By that I mean seeing how the larger sociopolitical structure not only influences and mirrors, but is also constituted in, the events and interactions in everyday classrooms.

For example, my ethnographic classroom research in Hungary (Duff 1993c, 1995, 1996, 1997) featured three diverse sites (schools) as cases in order to examine issues connected with an innovative dual-language (English immersion/bilingual) public education system: one in the capital city, one in a small resort town, and the last in a relatively distant provincial capital. By selecting heterogeneous contexts, in terms of geography, administration, human and other resources, student characteristics, and so on, it was possible to compare and contrast conditions, learning processes, and outcomes across the three instances of implementing the new model of education rather than take the findings from the most accessible and best-resourced site (in the capital city) as typical or representative. This cross-case comparison of different schools and multiple teachers and students within each school gave me a better understanding of what was or was not general in the implementation of dual-language education and in classroom discourse practices specifically and revealed how each school was dealing with sociopolitical and educational change as well as curricular change. It also necessitated analyzing the macroscopic social, political, historical, and educational contexts in which school reforms, including the introduction of English teaching and immersion and a more pro-Western curriculum and approach to teaching, were being introduced. At the same time, the local contexts and classroom discourse practices (e.g., in connection with recitation and assessment activities) were evidence of precisely the kinds of ambivalence, struggle and transformation that were witnessed at the national level and region, in the wake of Hungary's restored political autonomy, the dissolution of the USSR, and democratization in Central Europe.

Complexity

The theme of complexity in human behavior and interactions is foregrounded in an essay by Donmoyer (1990) on generalizability in case studies. He gives an account of a debate a few decades ago in psychology regarding whether all human behavior can be studied and described in terms of regularities in causes and *predictable* effects. According to Donmoyer, the eminent psychology scholar Cronbach, who had once advocated for ultimate generalizability, finally "concluded that human action is constructed, not caused, and that to expect Newton-like generalizations describing human action, as Thorndike did, is to engage in a process akin to 'waiting for Godot"' (Cronbach 1982, cited in Donmoyer 1990: 178). That is to say that the complex interactions that underlie human behavior need to be understood as co-constructed, unpredictable phenomena within society and within individuals. Capturing this complexity has long been a common theme and pursuit in qualitative research.

Larsen-Freeman (1997, 2002) has in recent years examined complexity in research in the natural and social sciences, generally, and in applied linguistics, in particular. Drawing on chaos and complexity theory (C/CT), a perspective originally developed to explain phenomena in the physical sciences, she has attempted to apply principles of C/CT to second language acquisition, to explain "mechanisms of acquisition, [the] definition of learning, the instability and stability of interlanguage, differential success, and the effect of instruction" (p. 152). Natural systems, she reports, are now understood by many scientists to be "dynamic, complex, nonlinear, chaotic, unpredictable, sensitive to initial conditions, open, self-organizing, feedback sensitive, and adaptive" (1997: 142). They change over time, usually consist of many component parts or agents that must be analyzed in light of the whole system, and are highly interactive, contingent, and interdependent, which is why complex systems are often unpredictable and their behaviors and properties are emergent (Larsen-Freeman 1997). Her own interests are principally grammar, language, and the processes of learning and using language. Qualitative research in applied linguistics is not restricted to these domains of course, although much second-language research (both quantitative and qualitative) focuses on these. If the diffusion of innovation were the object of inquiry (e.g., in language education curriculum reforms, in diachronic linguistic change, in the brain's functioning and organization, or in code-switching practices), then a complexity analysis would include the change processes, agents, and incremental (or dramatic, non-incremental) developments themselves, operating at the micro-level, in a bottom-up fashion, but viewed also within the context of macro-level or top-down influences.

Attempting to document highly complex social and linguistic phenomena in applied linguistics and also arrive at more general observations about trends, interrelationships among factors, and outcomes of interest to the field can be a difficult balancing act. A long, overly complicated story is difficult to tell and perhaps even more difficult for readers to transfer to other contexts or to relate to existing theories or models of learning, performance, or institutional change, more generally. In short, it becomes hard to see the proverbial forest for the trees and for the other elements and interactions foregrounded in the ecosystem. In case studies, ethnographies, and other types of mainstream qualitative research, thick description and triangulation, combined with data matrix displays, networks and cognitive maps that show the interaction and clustering among dimensions can help readers appreciate the complexity of the case(s) while at the same time situate the observations within a coherent and accessible conceptual framework (Miles & Huberman 1994).

Credibility

In my ethnographic classroom research in high school classrooms in Canada (Duff 2001, 2002b, 2004), I describe the changing demographics of urban areas and schools in Vancouver, as a result of immigration primarily and the impact of those changes on schooling. I tried to enhance the credibility of my findings regarding the experiences of immigrant non-native English speakers (NNESs) in mainstream content classes in the following ways: I selected two teachers, one male and one female, both deemed to be good models by their principal and peers. Both taught the same social studies course, one that is required for all Grade 10 students. The course, furthermore, is one of the first that newly mainstreamed NNESs must take. Both courses had a number of NNESs in them (one more than the other though) and the lessons were regularly observed and audio/video-recorded over an extended period of time (from six months to most of the academic year). Teachers and students (both NNESs and native speakers) were interviewed about their perceptions and experiences regarding students' participation in classroom discussions and in the school community, and about aspects of classroom discourse and content that proved most challenging. Other teachers and administrators in the school were also interviewed about these themes and relevant documents (e.g., an official accreditation report, students' writing, the course textbook) were consulted. Data analysis involved an ongoing examination of all the transcribed data for salient, recurring themes, examples of representative classroom interaction patterns and topics, and comparisons across the two classes for similarities and differences across themes. One sort of focal activity type was selected for anal-

ysis and comparison across lessons and courses: namely, teacher-fronted class discussions of current events and recently viewed educational films.

Missing from the data analysis to date, however, is a comprehensive treatment of the entire, rather vast data set, which is unwieldy for a single journal article or book chapter. Instead, I have conducted theme-driven analyses of illustrative excerpts or focal students, described with as much contextual information as possible, and selected precisely because of their typicality (which an analysis of the entire dataset and prolonged engagement in the field permits) and perceived theoretical import. Would another researcher draw the same inferences as I have from examining the same dataset – e.g., about the role and functions of pop culture in classroom discourse, or about the non-overt-participation of NNESs in mainstream courses? Not necessarily, but quite possibly they would. They would certainly note the silence of NNESs in class discussions from an analysis of their participation patterns, although their interpretations and explanations might differ. In fact, readers are able to draw their own conclusions if enough data is presented to them in the body of an article or appendix. Would others choose to focus on the same themes that I have selected and that emerged from the data? Again, not necessarily. The choice of themes related to integration and participation in discourse and about challenging aspects of academic discourse, the questions posed during interviews, and the focal speech events or activities singled out for analysis across lessons or courses might well have been different. However, in such analyses the authenticity, credibility, and trustworthiness of the analysis and report is also a product of the amount and type of data presented to support the findings and the provision of counter-examples, if any.

As Bromley (1986) argues, in the context of case study research,

> A particularly subtle source of error in case studies is the *absence* of information and ideas. The possibilities that no-one thought of, and the facts that were not known, must have invalidated numerous case-studies, simply because people's attention tends to be concentrated on the information actually presented. It is important in case-studies and in problem-solving generally to go "beyond the information given" to "what might be" the case. Naturally, such speculations need to be backed up by reasoned argument and by a search for relevant evidence. (p. 238)

There is always the possibility, then, that other crucial pieces of information – or alternate explanations – have been overlooked. One way of addressing this latter concern is by conducting not only a triangulation of data and methods, but also a triangulation of theory and researchers – that is, interpreting

the same observations or data from multiple theoretical or disciplinary standpoints or by inviting other researchers to analyze the same dataset (see e.g., Koschmann 1999, for an example of this in the analysis of discourse and interaction in a short medical tutorial). Naturally, in this kind of triangulation, datasets (e.g., transcripts, video clips) shared for analysis already represent a certain theoretical perspective and pre-selection, thus they cannot be construed as theory-neutral objects either.

Capturing participants' (or emic) perspectives to bolster credibility

One technique that is often discussed in qualitative research as a way of ensuring the authenticity or credibility of interpretations – and also a way of triangulating or verifying different perspectives and interpretations – is to conduct "member checks" and have the teachers or students involved read written reports before they are published and then incorporate their feedback or corrections; or to consult with them during the analysis. I have only done member checks informally in the past, by means of regular, sometimes fleeting conversations before or after classes; furthermore, my interaction with past research participants after data collection has ended has tended to be infrequent at best, simply because they or I have moved on to other endeavors and teachers and students tend to be very busy. In some of my studies, there has been some deliberate follow-up at a later point though (e.g., Duff, Wong, & Early 2000; Duff 2003) and in Hungary, I hired many of the students to transcribe classroom data for me, so I got their perspectives through discussions of the transcripts they had done.

There is great potential value in interviewing participants about their perspectives (either in their L1 or L2) before, during, or after a study, and especially minority students who, in my Canadian high school study, tended not to speak during class discussions and who otherwise provided little analyzable classroom data, apart from silence or nonverbal behaviors (Duff 2002b). Interviewing them provided insights into their English proficiency and their experiences both in and out of class. Obtaining teachers' perspectives about why they do what they do, and understanding their dilemmas, histories using, say, a certain teaching approach or task, or about their perspectives on students' comprehension, performance, and so on can yield rich insights (Duff & Li 2004). In short, the triangulation of research *methods*, *data*, and participant *perspectives* (including my own analytic perspectives) to shed light on classroom phenomena provides a more multidimensional, richer image or composite and thus sys-

temic understanding than an individual snapshot or series of snapshots taken from a distance would.

However, obtaining participants' perspectives may be appropriate for some qualitative studies or for certain kinds analyses (e.g., task-based behavior) but of limited usefulness when analyzing learners' metalinguistic awareness or use of particular forms (e.g., asking Mandarin speakers about their use of a particular aspect marker, *le,* or asking people about their L1-L2 code-switching), depending on participants' educational background, proficiency, age, and self-awareness. The ability to ascertain participants' perspectives depends very much on their L2 competence and ability to report on things in their L2, unless the researcher understands their L1 or an interpreter is present. It also depends on participants' metalinguistic or metacognitive awareness and, in the end, their accounts are, like many other forms of data, undeniably social and narrative constructions.

Enhancing generalizability: Representativeness or typicality of cases selected

One of the key ways in which case studies and ethnographies attempt to approach generalizability, if indeed that is sought – or at least to avoid the criticism of unique or idiosyncratic findings whose impact on knowledge more broadly may be difficult to assess or assert – is to choose typical or representative sites or participants to study and not just those that are most convenient or easily accessible and whose representativeness is unclear. This is not a goal shared by all qualitative researchers though, who may choose to study people and sites already known to them or to which they have access; most accessible, of course, are the researchers' themselves through introspective (e.g., diary/narrative) studies, memoirs, and auto-ethnographies (Schumann 1997).

Here I provide an example of case selection from my own case study research and the issue of typicality. I loosely modeled my longitudinal SLA study of a Cambodian immigrant to Canada learning English (Duff 1993a) after an influential longitudinal study conducted by Huebner (1983) of a Hmong-speaking Laotian immigrant to Hawaii. I examined syntactic features in the interlanguage of my research participant that had parallels in Huebner's study. Both cases were refugees of a similar age and social class who came from somewhat similar, topic-comment Southeast Asian languages. And, while similarities emerged in how they expressed existentials in English, based on their first language structures (e.g., *in camp have/has many soldier* instead of *there*

were many soldiers in the camp), in neither case were the learners selected because of being typical, let alone "prototypical," of learners from their source communities. To do so, would have required some prior or subsequent sampling from learners in the source communities, or a concurrent cross-sectional study involving other Cambodians or Laotians or Hmong speakers, some kind of census data about the communities involved, or a multiple case study addressing the issue of representativeness or variation. Rather, both subjects were samples of convenience – willing, personable, and available research participants whose English was developmentally intriguing.

Thus, it would be impossible to claim that other (or *all*) Hmong or Cambodian learners would have proceeded in English with exactly the same strategies for topic marking or existential constructions, for example, but the two studies taken together, plus some interlanguage studies of Chinese and Japanese students who proceeded similarly (in part because of their similarly configured topic-comment L1s), suggested some patterns or potential repertoires that learners from those backgrounds are likely to avail themselves of (Duff 1993a). Only with additional cases or supplementary studies (not necessarily longitudinal or as in-depth) with learners from the same backgrounds would it be possible to assert that specific interlanguage structures (e.g., the use of *have* vs. *has* as the default generic existential verb; or the use of *isa* as a topic marker) were shared by all Cambodian or Hmong ESL learners. The spirit of the "performance analysis" SLA studies of the day, moreover, was to see what learners were capable of doing, sometimes in highly creative and unique ways, with the linguistic and cognitive resources at their disposal, and not just how they fell short of target norms (Larsen-Freeman & Long 1991).

Furthermore, the typicality of a case within one context (e.g., a Cambodian learner in Canada) may not equate with typicality in another (e.g., a Cambodian learner in Cambodia), and not all case study or ethnographic work in our field sets out to deal with typical or "normal" learners, teachers, or programs, in any event. Some are targeted for research precisely because of their atypicality, uniqueness, resilience, or even pathology in some instances; they are purposefully and opportunistically selected because of the insights they are expected to generate about the possibilities of language learning and use. (See Peräkylä (1997) for a discussion of "possibilities" in connection with generalizability in conversation analysis.) For example, Genie was a famous case discussed in both first and second language acquisition from the late 1970s (Curtiss 1977), when her situation of abject neglect was first discovered and a team of researchers set out to study and help her. Genie, who had been deprived of normal human language and interaction for most of the first 13 years of her life, was seen to

be a test case for the critical period hypothesis: evidence that she could learn language (English, her L1) successfully even after the onset of puberty, it was asserted, might discredit the critical period hypothesis for language acquisition. In the end, Genie's language development (e.g., morphology and syntax) was quite modest and her lack of targetlike proficiency in English was difficult to explain, precisely because of her extreme atypicality, but one explanation given was the existence of a critical period for language acquisition – which she had passed. The inferences drawn about Genie in connection with this hypothesis were perhaps not completely warranted, especially considering her highly atypical social-psychological history, which included far more than just the presence or absence of language learning opportunities, but also deprived her of normal human attachments, basic opportunities for cognitive stimulation and development, and so on. However, had Genie been able to achieve something approximating "normal" native proficiency in English despite her early deprivation, the inferences and generalizations drawn from the study would have been far more powerful and also remarkable.

Other atypical cases and situations that have been explored include highly successful (or "talented") and highly unsuccessful language learners (e.g., Ioup 1989; Ioup, Boustagui, El Tigi, & Moselle 1994). The reason for selecting cases along a continuum of experience, whether extreme cases, critical cases, or typical cases, is to explore the range of human (linguistic) possibilities along a particular dimension. Yet despite researchers' frequent disclaimers and cautionary notes about the generalizability of their case studies, the theoretical findings from seminal early case studies in SLA (e.g., those in the 1970s and 1980s by Schumann, Schmidt, Hatch and colleagues; e.g., Hatch 1978), as well as more recent influential studies examining gender, ethnicity, and identity in SLA (e.g., Norton 2000; Norton & Toohey 2001), have nevertheless achieved fairly wide generalization within the field, giving rise to important new understandings of second language learning. For example, generalizations have been made about the Acculturation Model (for and against) in relation to SLA, the impact of noticing gaps on SLA, fossilization, language loss, and the role of identity, power, and motivation/investment in SLA. In fact, some of the most influential generalizations and models of SLA have originated from just a small handful of case studies.

In fairness to the work and its legacy, this kind of theoretical (over) generalization from cases sampled by convenience has not only occurred in hallmark qualitative studies. Because of the lack of replication in many kinds of quantitative SLA or classroom-oriented studies as well, findings from small (e.g. quasi-experimental or otherwise) one-off studies or even larger studies

conducted with a particular population, are similarly taken as proof that, for example, gender (or task familiarity, partner familiarity, ethnicity, same-L1 status, etc.) does or does not make a difference in language learning or use, that learners do not generally learn from one another's errors, and so on; or, in more abstract terms, that variable-x has such-and-such effect on variable- or population-y (or, when generalized, to possibly all language learners). In part, this extension of findings in the absence of widespread evidence is a symptom of a young field that may seek originality or novelty in studies, more so than the robustness, durability, or replicability of findings with different pools of subjects/participants and contexts, on which basis they can support theoretical conclusions previously drawn.

Another method for enhancing the potential generalizability as well as credibility of qualitative research is to conduct multisite or multiple-case studies, not all of which may have the same attributes. Schofield (1990) asserts that "a finding emerging from the study of several very heterogeneous sites would be more robust and thus more likely to be useful in understanding various other sites than one emerging from the study of several very similar sites" (p. 212). We could also replace "sites" with "people." If we found similar developmental trends across three Cambodian learners – the original subject referred to above (a refugee with interrupted education), an instructed university student, and a child learner – it would add to the robustness of the original observations or suggest alternate developmental pathways. On this same theme, in lieu of multiple cases or larger surveys of general trends, researchers may look for an aggregation of case studies (comparable to a meta-analysis of quantitative studies) to corroborate findings across studies. Stake (2000) refers to these as *collective* case studies (of either a similar or different nature). Researchers may also include participants' own judgments about generalizability or representativeness to assist them with interpretations about typicality (Hammersley 1992). Finally, longitudinal studies with data collected from multiple sources or task types also have the potential to increase the nature of inferences that can be drawn about learning because the developmental pathways (consistent, incremental, erratic, or very dynamic) can be shown, as well as interactions in the acquisition of several interrelated structures over time (Huebner 1983).

The approach taken by Harklau (1994), Leki (1995), McKay and Wong (1996), Toohey (2000), and Willett (1995) in their longitudinal ethnographic classroom-based or classroom-oriented studies in TESOL/applied linguistics was to select three to five cases for in-depth analysis within the context of classrooms with diverse (e.g., immigrant, minority, local) learners (see Duff in press, for details). Sometimes focal cases in qualitative research are also

analyzed against the backdrop of a larger set of participants and data (e.g., Kouritzen 1999; Li 1998). The authors may marshal various kinds of evidence (e.g., excerpts from interviews or classroom interactions) to support their claims. For example, Harklau's article included 21 short excerpts from interview data taken primarily from students to support her observations, organized around the themes of spoken language use in the [mainstream] classroom, spoken language use in the ESL classroom, written language use in the mainstream, written language use in the ESL classroom, structure and goals of instruction, explicit language instruction, and socializing functions of schooling. Leki (1995), on the other hand, focused on the challenges faced by three graduate and two undergraduate international (visa) students from Europe and Asia in their first semester at an American university. Of interest was the English writing requirements in their disciplinary courses across the curriculum and their coping strategies, as newcomers to the local academic culture. The data presented include well-rounded profiles of five focal students (the cases), followed by a description and discussion of ten general themes (strategies) that surfaced across the five students' experiences, as well as differences among them. Nine short quotations or excerpts from the students' interviews, journals, or assignments were included from the corpus of transcribed data. Neither study included excerpts of classroom discourse.

Most of my graduate students conducting case studies or ethnographies similarly select 4–6 focal participants for study in one or more sites (Duff & Uchida 1997; Morita 2002, 2004; Kobayashi 2003). Choosing 6 initially means that if there is attrition among participants, there will likely be 3–4 cases remaining, providing multiple exemplars of the phenomenon under investigation. Almost all have collected data through extensive and sustained classroom observations, interviews, document analysis, artifacts (e.g., term papers, PowerPoint presentations), researcher fieldnotes or journals, and sometimes participants' journals or logs as well. The greater the number of participants, cases or sites, however, the less possible it is to provide an in-depth description and contextualization of each one, taking fully into account the complexity of interactions, the perspectives of the participants, and so on. However, having more than one focal case can provide interesting contrasts or corroboration across cases. The cases are carefully selected from the larger set of potential candidates, based on oral proficiency scores, gender, age, length of time in a program, university major, or other such variables deemed important in relation to the research questions. Exceptional cases, just like counter-examples to findings, need to be explained.

Although Polio and Duff (1994) was neither an ethnography nor a case study in the core sense of being in-depth analysis of one context, by conducting a qualitative analysis of instructional language practices (in terms of L1 vs. L2 use) across first year college-level courses in 13 different languages (rather than just one or two), many of which were typologically unrelated (e.g., Spanish, Korean, Polish, Hebrew), we were able to determine both the range of L1 use in L2 teaching at one particular university, and also some similarities across the courses with regard to the functions of L1 use regardless of the L2. Although we couldn't generalize the descriptive statistics about L1-L2 use from this university, the heterogeneity of the sample provided a good foundation for the possible transferability of the findings to other universities or sites – and certainly raised questions for future research as to (1) how much L1 vs. L2 is used in foreign language classrooms; (2) what the functions of L1 vs. L2 are in classroom discourse; and (3) how to increase L2 use – which we argued was a worthy and theoretically and pedagogically defensible goal. Subsequently, certain other researchers have questioned the premise (that more L2 use is better; e.g., Cook 2001) but that that has not hampered the transferability of the findings regarding pervasive L1 use in certain foreign language courses and reasons for that.

In this section, I have discussed ways of trying to establish typicality, representativeness, or even maximum variation across cases, although many researchers choose their case study or ethnography sites and participants more opportunistically. Stake (1995, 2000), a case study methodologist, differentiates between what he calls *intrinsic* and *instrumental* case studies (categories that he concedes are best seen on a continuum) and does not insist on seeking representativeness in all case studies. He also claims that studies differ in the claims they may wish to make regarding generalizability:

> I call a study an *intrinsic case study* if it is undertaken because, first and last, the researcher wants better understanding of this particular case. Here, it is not undertaken primarily because the case represents other cases or because it illustrates a particular trait or problem, but because, in all its peculiarity *and* ordinariness, this case itself is of interest. … The purpose is not to come to understand some abstract construct or generic phenomenon, such as literacy or teenage drug use or what a school principal does. The purpose is not theory building – although at other times the researcher may do just that.
>
> … I call it *instrumental case study* if a particular case is examined mainly to provide insight into an issue or to draw a generalization. The case is of secondary interest, it plays a supporting role, and it facilitates our understanding of something else. The case is still looked at in depth, its contexts scrutinized,

> its ordinary activities detailed, but all because this helps the researcher to pursue the external interest. The case may be seen as typical of other cases or not... (Stake 2000: 437)

The notion of instrumental case study is related to the concept of analytic generalization (generalization to theory not to populations), which I describe in the next section. Stake also suggests that "[e]ven intrinsic case study can be seen as a small step toward grand generalization..., especially in the case that runs counter to the existing rule" (p. 439). A good example of this is Schmidt's (1983) analysis of Wes, a Japanese artist living in Hawaii whose English did not develop very well despite his high degree of acculturation in the local English-speaking American community. I do not believe that this research subject was selected for his intrinsic interest value alone but, regardless, Schmidt used this well known case as a way of refuting (if not "falsifying") Schumann's Acculturation Model; the model, based largely on Schumann's case study of a Costa Rican immigrant to America named "Alberto," posited that acculturation is a major causal factor in successful SLA (Schumann 1978).

Analytic generalization

My SLE research has examined the linguistic and/or interactional behavior of students in a variety of types of classrooms. Increasingly, I have framed the work in terms of language socialization involving the ethnography of communication and/or case studies, and have added poststructural interpretations of data in some cases, although the exact combination of methods and the thematic focus depends on the research questions being addressed (Duff 2002a). Studying a similar theoretical process (language socialization) across very different contexts (in dual-language schools in Hungary, in heterogeneous Canadian high schools and vocational settings) provides a level of generalization as well, about the construct or process in question (Duff 2003). The generalizability, then, is not to populations but to theoretical models, often captured in simple diagrams, which also take into account the complexity of second language learning and the multiple possible outcomes that exist. However, Bromley (1986) notes that even producing diagrams capturing processes, which is a form of data or theme reduction as well as representation, provides a level of abstraction and generality beyond the details of the local case(s) (see also Miles & Huberman 1994, for illustrations of the multiple ways in which data can be visually presented). Such models and heuristics must be backed up

with logical reasoning and evidence that warrant the inferences that are drawn by the researcher or may be drawn by others (Bromley 1986).

Conclusion

In this chapter, I have presented some basic contrasts between the way the terms inference and generalizability are operationalized – or contested – in quantitative research and qualitative research, underscoring the "multiple research perspectives" promised in the subtitle of this book. I noted that many qualitative researchers have rejected generalizability in the traditional sense as an achievable or desirable objective of research, although others with a more positivist bent are more committed to it (e.g., Yin 2003; Miles & Huberman 1994). Qualitative researchers embracing generalizability suggest that careful sampling for representativeness, conducting multisite and multiple case studies, providing careful chains of reasoning (inference and interpretation), and explaining aberrant data or cases enhances the generality of the findings of studies. On the other hand, those rejecting traditional notions of generalizability highlight elements more associated with internal validity and reliability (infrequently using those terms), and effective reasoning and writing instead: for example, by providing thick description, credible evidence, thorough data analysis, appropriate representation of contexts and data so that readers can learn from others' experiences and draw their own conclusions (or inferences) about transferability and relevance. I then provided examples of how my own and others' case studies and ethnographic research have either broached generalizability (and) or instead sought to foreground contextualization, complexity and credibility of studies of language education and learning.

A final comment relates to reporting in applied linguistics research. No matter how generalizable a study might be, the results of research are often inaccessible or incomprehensible to others because of the discourse associated with the inquiry tradition. If readers are unable to easily comprehend the results and understand the chains of reasoning, the new knowledge or insights will not "travel" or be transferable to new settings. Instead, the findings may only resonate with a small number of like-minded researchers and thus may have less impact on the field than intended or deserved. One of the potential benefits of qualitative research on its own or in combination with quantitative studies is that it can provide concrete, situated instances of an abstract phenomenon, which, when done well, may contribute meaningfully to theory-building and to knowledge in the field. Qualitative research has its

own share of challenges (and diversity, as Lazaraton 2003, argues in her comparison of ethnography and conversation analysis), foremost perhaps being the difficulty of coming up with a powerful, elegant and relatively straightforward description of the complex relationships among many factors observed in cases or sites of interest, while at the same time demonstrating that the interpretations are solidly grounded in empirical data and observation. As new approaches to research in applied linguistics emerge and evolve, the criteria, genres, and possibilities for reporting and evaluating research will undoubtedly change as well and new perspectives on inference and generalizability will result (Duff 2002a). In addition, ideally more multi-method, multi-site and multinational/multilingual studies of SL acquisition and education phenomena will be conducted to demonstrate the generality of the conditions and findings of research in critical areas.

Notes

1. In testing, there are particular statistical procedures connected with an area known as *Generalizability Theory* as well.

2. Another nagging problem with many small-scale quantitative studies in SLA/SLE is that our means of quantifying and drawing statistical inferences is often less than ideal. There is a lack of robust, truly appropriate statistical procedures for the kinds of comparisons of frequency/nominal data typical in discourse analyses (with repeated measures) of the sort often used in task-based research, for example. As Hatch and Lazaraton (1991) and Lazaraton (2000) point out, in various analyses of linguistic production and interaction, multiple t-tests, ANOVAs, chi-square statistics, and factor analyses are often used inappropriately; assumptions underlining the statistics may not be met or tests are used in ways for which they weren't intended mathematically – e.g., by conducting multiple t-tests, by using powerful parametric tests with nonparametric data, or using nonparametric tests with the wrong kind of nonparametric data, especially for within-subject comparisons of nominal data, such as in comparisons of linguistic constructions of one type produced by research participants compared with other types of constructions the same people use when performing one or more different tasks. All of these difficulties may inadvertently compromise the inferences and generalization claimed by researchers.

References

Adler, P. A. & Adler, P. (1994). Observational techniques. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 377–392). Thousand Oaks, CA: Sage.

Altheide, D. L. & Johnson, J. M. (1994). Criteria for assessing interpretive validity in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 485–499). Thousand Oaks, CA: Sage.

Bogdan, R. C. & Biklen, S. K. (1992). *Qualitative research for education* (2nd ed.). Boston: Allyn and Bacon.

Bracht, G. H. & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5*, 437–478.

Breen, M. (1985). The social context for language learning – A neglected situation? *Studies in Second Language Acquisition, 7*, 135–158.

Bromley, D. B. (1986). *The case-study method in psychology and related disciplines*. New York, NY: John Wiley & Sons.

Brown, J. D. & Rogers, T. S. (2002). *Doing second language research*. Oxford: OUP.

Chapelle, C. & Duff, P. (Eds.). (2003). Some guidelines for conducting quantitative and qualitative research in TESOL. *TESOL Quarterly, 37*, 157–178.

Cook, V. (2001). Using the first language in the classroom. *Canadian Modern Language Review, 57,* 402–423.

Coughlan, P. & Duff, P. (1994). Same task, different activities: Analysis of a SLA [second language acquisition] task from an Activity Theory perspective. In J. Lantolf & G. Appel (Eds.), *Vygotskian perspectives on second language research* (pp. 173–193). New Jersey: Ablex.

Crabtree, B. & Miller, W. (Eds.). (1999). *Doing qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Curtiss, S. (1977). *Genie: A psycholinguistic study of a modern-day "wild child."* New York: Academic Press.

Davis, K. (1995). Qualitative theory and methods in applied linguistics research. *TESOL Quarterly, 29*, 427–453.

Denzin, N. & Lincoln, Y. (Eds.). (2000). *Handbook of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Donmoyer, R. (1990). Generalizability and the single-case study. In E. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 201–232). New York, NY: Teachers College Press.

Duff, P. (1993a). Syntax, semantics, and SLA: The convergence of possessive and existential constructions. *Studies in Second Language Acquisition, 15,* 1–34.

Duff, P. (1993b). Tasks and interlanguage performance: An SLA [second language acquisition] research perspective. In G. Crookes & S. Gass (Eds.), *Tasks in language learning: Integrating theory and practice* (pp. 57–95). Clevedon, UK: Multilingual Matters.

Duff, P. (1993c). Changing times, changing minds: Language socialization in Hungarian-English schools. PhD Dissertation, University of California, Los Angeles.

Duff, P. (1995). An ethnography of communication in immersion classrooms in Hungary. *TESOL Quarterly, 29,* 505–537.

Duff, P. (1996). Different languages, different practices: Socialization of discourse competence in dual-language school classrooms in Hungary. In K. Bailey & D. Nunan (Eds.), *Voices from the language classroom: Qualitative research in second language acquisition* (pp. 407–433). New York: CUP.

Duff, P. (1997). Immersion in Hungary: An EFL experiment. In R. K. Johnson & M. Swain (Eds.), *Immersion education: International perspectives* (pp. 19–43). New York, NY: CUP.

Duff, P. (2001). Language, literacy, content and (pop) culture: Challenges for ESL students in mainstream courses. *Canadian Modern Language Review, 58,* 103–132.

Duff, P. (2002a). Research approaches in applied linguistics. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 13–23). Oxford: OUP.

Duff, P. (2002b). The discursive co-construction of knowledge, identity, and difference: An ethnography of communication in the high school mainstream. *Applied Linguistics, 23*, 289–322.

Duff, P. (2003, March). New directions and issues in second language socialization research. Plenary talk presented at the annual meeting of the *American Association for Applied Linguistics*, Arlington, Virginia.

Duff, P. (2004). Intertextuality and hybrid discourses: The infusion of pop culture in educational discourse. *Linguistics and Education*, *14*(3–4), 231–276.

Duff, P. (in press). Qualitative approaches to second language classroom research. In J. Cummins & C. Davison (Eds.), *Handbook of English language teaching*. Dordrecht: Kluwer.

Duff, P. & Early, M. (1996). Problematics of classroom research across sociopolitical contexts. In S. Gass & J. Schachter (Eds.), *Second language classroom research: Issues and opportunities* (pp. 1–30). Hillsdale, NJ: Lawrence Erlbaum.

Duff, P. & Li, D. (2004). Issues in Mandarin language instruction: Theory, research, and practice. *System*, *32*, 443–456.

Duff, P. & Uchida, Y. (1997). The negotiation of teachers' sociocultural identities and practices in postsecondary EFL classrooms. *TESOL Quarterly, 31*, 451–486.

Duff, P., Wong, P., & Early, M. (2000). Learning language for work and life: The linguistic socialization of immigrant Canadians seeking careers in healthcare. *Canadian Modern Language Review, 57,* 9–57.

Duranti, A. & Goodwin, C. (Eds.). (1992). *Rethinking context: Language as an interactive phenomenon*. Cambridge: CUP.

Edge, J. & Richards, K. (1998). May I see your warrant, please?: Justifying outcomes in qualitative research. *Applied Linguistics, 19*, 334–356.

Eisner, E. & Peshkin, A. (1990). *Qualitative inquiry in education: The continuing debate.* New York, NY: Teachers College Press.

Firestone, W. A. (1993). Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher, 22*, 16–23.

Fraenkel, J. R. & Wallen, N. E. (1996). *How to design and evaluate research in education* (3rd ed.). New York: McGraw Hill.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research* (6th ed.). White Plains, NY: Longman.

Gall, M. D., Gall, J. P., & Borg, W. T. (2003). *Educational research* (7th ed.). White Plains, NY: Pearson Education.

Hammersley, M. (1992). Some reflections on ethnography and validity. *Qualitative Studies in Education, 5,* 195–204.
Harklau, L. (1994). ESL versus mainstream classes: Contrasting L2 learning environments. *TESOL Quarterly, 28*, 241–272.
Hatch, E. (Ed.). (1978). *Second language acquisition*. Rowley, MA: Newbury House.
Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston, MA: Heinle and Heinle.
Holliday, A. (1994). *Appropriate methodology and social context*. Cambridge: CUP.
Holliday, A. (2002). *Doing and writing qualitative research*. Thousand Oaks, CA: Sage.
Huebner, T. (1983). *A longitudinal analysis of the acquisition of English*. Ann Arbor, MI: Karoma Publishers.
Hulstijn, J. & DeKeyser, R. (Eds.). (1997). *Testing SLA theory in the research laboratory*. Special issue, *Studies in Second Language Acquisition, 19*, 2.
Ioup, G. (1989). Immigrant children who have failed to acquire native English. In S. Gass, C. Madden, D. Preston, & L. Selinker (Eds.), *Variation in second language acquisition: Psycholinguistic issues* (pp. 160–175). Clevedon, UK: Multilingual Matters.
Ioup, G., Boustagui, E., El Tigi, M., & Moselle, M. (1994). Re-examining the critical period hypothesis: A case study of successful adult second language acquisition in a naturalistic environment. *Studies in Second Language Acquisition, 16*, 73–98.
Johnson, D. M. (1992). *Approaches to research in second language learning*. New York, NY: Longman.
Kobayashi, M. (2003). The role of peer support in ESL students' accomplishment of oral academic tasks. *Canadian Modern Language Review, 59,* 337–368.
Kouritzen, S. (1999). *Face[t]s of first language loss*. Mahwah, NJ: Lawrence Erlbaum.
Koschmann, T. (1999). Meaning making. Special issue of *Discourse Processes, 27*(2).
Krathwohl, D. (1993). *Methods of educational and social science research*. White Plains, NY: Longman.
Kramsch, C. (Ed.). (2002). *Language acquisition and language socialization: Ecological perspectives.* New York, NY: Continuum.
Larsen-Freeman, L. (2002). Language acquisition and use from a chaos/complexity theory perspective. In C. Kramsch (Ed.), *Language acquisition and language socialization* (pp. 33–46). New York, NY: Continuum.
Larsen-Freeman, D. & Long, M. H. (1991). *An introduction to second acquisition research*. New York, NY: Longman.
Lather, P. (1991). *Getting smart: Feminist research and pedagogy within the postmodern*. New York: Routledge.
Lazaraton, A. (2000). Current trends in research methodology and statistics in applied linguistics. *TESOL Quarterly, 34,* 175–181.
Lazaraton, A. (2003). Evaluating criteria for qualitative research in applied linguistics: Whose criteria and whose research? *Modern Language Journal, 87*, 1–12.
Leki, I. (1995). Coping strategies of ESL students in writing tasks across the curriculum. *TESOL Quarterly, 29*, 235–260.
Li, D. (1998). Expressing needs and wants in a second language: An ethnographic study of Chinese immigrant women's requesting behavior. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Lincoln, Y. & Guba, E. G. (1985). *Naturalistic inquiry.* Beverly Hills, CA: Sage.
Maxwell, J. A. (1996). *Qualitative research design: An interactive approach*. Thousand Oaks, CA: Sage.
McKay, S. L. & Wong, S. C. (1996). Multiple discourses, multiple identities: Investment and agency in second-language learning among Chinese adolescent immigrant students. *Harvard Educational Review, 66*, 577–608.
Merriam, S. (1998). *Qualitative research and case study applications in education* (2nd ed.). San Francisco, CA: Jossey-Bass.
Miles, M. & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.
Morita, N. (2002). Negotiating participation in second language academic communities: A study of identity, agency, and transformation. PhD Dissertation, University of British Columbia.
Morita, N. (2004). Negotiating participation and identity in second language academic communities. *TESOL Quarterly, 38*, 573–603.
Neuman, W. L. (1994). *Social research methods: Qualitative and quantitative approaches* (2nd ed.). Boston: Allyn & Bacon.
Norton, B. (2000). *Identity and language learning: Gender, ethnicity and educational change*. London: Pearson Education.
Norton, B. & Toohey, K. (2001). Changing perspectives on good language learners. *TESOL Quarterly, 35,* 307–322.
Palys, T. (1997). *Research decisions: Quantitative and qualitative perspectives* (2nd ed.). Toronto: Harcourt, Brace, Jovanovich.
Peräkylä, A. (1997). Reliability and validity in research based on tapes and transcripts. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice*. Thousand Oaks, CA: Sage.
Polio, C. & Duff, P. (1994). Teachers' language use in university foreign language classrooms: A qualitative analysis of English and target language alternation. *Modern Language Journal, 78,* 313–326.
Richardson, L. (1994). Writing: A method of inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 516–529). Thousand Oaks, CA: Sage.
Schmidt, R. (1983). Interaction, acculturation and the acquisition of communicative competence. In N. Wolfson & E. Judd (Eds.), *Sociolinguistics and language acquisition* (pp. 137–174). Rowley, MA: Newbury House.
Schmidt, R. & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 237–326). Rowley, MA: Newbury House.
Schofield, J. W. (1990). Increasing the generalizability of qualitative research. In E. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 201–232). New York, NY: Teachers College Press.
Schumann, J. (1978). *The pidginization process: A model for second language acquisition.* Rowley, MA: Newbury House.
Schumann, J. (1997). *The neurobiology of affect in language.* Malden, MA: Blackwell.
Silverman, D. (2000). *Doing qualitative research*. Thousand Oaks, CA: Sage.

Stake, R. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.
Stake, R. (2000). Case studies. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 435–454). Thousand Oaks, CA: Sage.
Toohey, K. (2000). *Learning English at school: Identity, social relations and classroom practice.* Clevedon, UK: Multilingual Matters.
van Lier, L. (1988). *The classroom and the language learner*. New York, NY: Longman.
van Lier, L. (1997). Observation from an ecological perspective. *TESOL Quarterly, 22*, 783–787.
Watson-Gegeo, K. (1988). Ethnography in ESL: Defining the essentials. *TESOL Quarterly, 22,* 575–592.
Willett, J. (1995). Becoming first graders in an L2: An ethnographic study of language socialization. *TESOL Quarterly, 29*, 473–504.
Yin, R. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: Sage.

Verbal protocols

What does it mean for research to use speaking as a data collection tool?*

Merrill Swain
The Ontario Institute for Studies of The University of Toronto

What do verbal protocols represent? The response to this question differs depending on the theoretical perspective one takes. I will examine the answer to this question from an information processing perspective and from a sociocultural theory of mind perspective. The different assumptions underlying the two theories lead to different interpretations of verbal protocol data. Research evidence is provided that suggests that verbal protocols, particularly stimulated recalls, are a source of learning. This suggests that verbal protocols cannot be used neutrally as a method of collecting data in second language acquisition research, but instead they need to be considered as part of the "treatment" when making claims about learning and development. Verbal protocols are not just "brain dumps"; rather they are a process of comprehending and reshaping experience – they are part of what constitutes development and learning.

In psychological research, we are interested in determining the processes individuals use when they engage in an activity. One source of information about those processes is the individual him/herself – who tells the researcher about what he or she is thinking during or after the activity. The data generated by this methodology are referred to as verbal protocols. The purpose of any data collection is to draw inferences from the data collected. Inference, in the context of the present chapter, deals with the type of interpretation that researchers can make with regard to verbal protocol data. In this chapter, I intend to discuss the different inferences that are drawn by information processing theorists and sociocultural (theory of mind) theorists from verbal protocols. My intent is to question inferences that are made within the more traditional cognitive paradigm, arguing for alternative interpretations of verbal protocol

data based on sociocultural research and theory. I also consider the extent to which we can generalize from inferences drawn from local and situated contexts (Chaloub-Deville 2003). The ultimate purpose of this chapter is to argue that we, as researchers, can no longer think of verbal protocols as a "neutral" methodology – that is, as a methodology that has no impact on our findings.

Gass and Mackey (2000) define verbal protocols as the data one gets "by asking individuals to vocalize what is going through their minds as they are solving a problem or performing a task" (p. 13). Verbal protocols can take the form of concurrent "think alouds", where individuals say what is going through their mind while they are in the process of solving the problem or performing the task (e.g., Cumming 1990). Or, verbal protocols can take the form of some sort of retrospective introspection, for example, a stimulated recall (e.g., Mackey 2002; Swain & Lapkin 2002). In a stimulated recall, individuals are provided with a stimulus which constitutes a bit of their past behaviour. For example, individuals may be shown a clip of a video in which they appear and are asked to talk about what was going through their minds at that particular time (e.g., Swain & Lapkin in press).

The specific question I wish to address in this chapter is: (1) *What do these verbal protocols represent?* This is followed by a section on relevant research. The implications for research on the different interpretations of verbalizing (speaking) are considered at the end of the chapter. Overall, I attempt to answer what it means for researchers in second language learning to use speaking as a data collection tool.

What do verbal protocols represent?

I examine this question from the perspective of two different theories of human cognition: information processing theory (Ericsson & Simon 1993, 1998) as representative of recent (last three decades) thinking in cognitive science, and a sociocultural theory of mind (e.g., Lantolf 2000; Vygotsky 1978; Wertsch 1985, 1991; Wertsch & Tulviste 1996). The issue which underlies the debate is no less than the relationship between language and thought – an issue that "has been little discussed in recent decades, since many have thought the issue to be closed" (Carruthers & Boucher 1998b: 2) because of the dominance of the computer metaphor for the mind in the cognitive sciences (see e.g., Searle 2002 for a critique of the metaphor). In this view, language is "but an input and output module for central cognition" (Carruthers & Boucher 1998b: 2). This view, however, is being challenged, even within the cognitive sciences as indicated

by a recent book edited by Carruthers and Boucher (1998a) "*Language and Thought: Interdisciplinary Themes.*" Also, this view is being challenged by those influenced by Vygotsky (1978) and his colleagues and students (e.g., Gal'perin 1969; Luria 1973), whose writings have only reached North American *second* language theoreticians and researchers in the last decade or so (e.g., Lantolf 2003; Lantolf & Appel 1994).

Carruthers and Boucher (1998b) suggest that there are roughly two opposing camps among those who are interested in the place of language in cognition. There are those who see "the exclusive function and purpose of language to be the communication of thought, where thought itself is largely independent of the means of its transmission from mind to mind" (p. 1). Information processing theory falls into this camp. Alternatively, there are those who see language as "crucially implicated in human thinking. . .that language itself is *constitutively involved in* [some kinds of thinking]" (p. 1). Language is not simply a vehicle for communication, but plays critical roles in creating, transforming, and augmenting higher mental processes. Sociocultural theory of mind falls into this second camp. The different assumptions underlying the two perspectives – information processing theory and sociocultural theory – lead to different interpretations of the data elicited in verbal protocols (Smagorinsky 1998).

According to information processing theory (Ericsson & Simon 1993), think alouds are a ***report*** of the (oral) contents of short-term memory, and represent a trace of the cognitive processes that people attend to while doing a task. In a stimulated recall, as Gass and Mackey (2000) point out, "the use of and access to memory structures is enhanced, if not guaranteed, by a prompt that aids in the recall of information" (p. 17). In both cases the assumption is that verbal protocols provide data for investigating cognition ***direct*** from memory. According to Ericsson and Simon (1993: 222), the verbalization "is a direct encoding of the heeded thought and reflects its structure". The data tell us "what information [individuals] are attending to while performing their tasks, and by revealing this information, can provide an orderly picture of the exact way in which the tasks are being performed: the strategies employed, the inferences drawn from information,. . ." (p. 220). In this way, verbal protocols provide the evidence from which models of human cognitive processing are generated.

Ericsson and Simon (1993) make a distinction between instructions to participants to verbalize thoughts *per se*, what they refer to as Type 1 and Type 2 verbalization, and instructions to verbalize specific information, such as reasons and explanations (Type 3 verbalization). Type 1 and Type 2 verbalizations

do not, they claim, change the sequencing of the cognitive processes, but the time to carry them out may be longer as a result of the verbalizing. As for Type 3 verbalization, they summarize their review of the research literature as: "...directing subjects to engage in specific thought activities with associated overt verbalization changes the cognitive processes and thus alters concurrent and retrospective performance" (p. xix). They continue, that "...the effects of directing verbalization do not involve any magical influences but can be understood in terms of the changes induced in the associated cognitive process *by the instructions*" (p. xix) (*my italics*). In other words, Ericsson and Simon understand the instructions as causal, not the verbalization.

> ...in the review of studies comparing different instructions to verbalize, we found substantial evidence that differences in performance were induced by telling the subject *how* to verbalize. In order to verbalize the information called for by the instructions, instead of the information he would normally have attended to, he had to change his thought processes.
>
> (Ericsson & Simon 1993: 107)

As Vygotsky (1986: 218) asked, however, "Does language only reflect thought (memory) or can it change thought (memory)?" Vygotsky believed that "thought is not merely expressed in words; it comes into existence through them" and that "thought undergoes many changes as it turns into speech: it finds its reality and form" (p. 219). "The process of rendering thinking into speech is not simply a matter of memory retrieval, but a process through which thinking reaches a new level of articulation" (Smagorinsky 1998: 172–173). Ideas are crystallized and sharpened, and inconsistencies become more obvious. Smagorinsky (2001) makes clear the implication of this position: "If thinking becomes rearticulated through the process of speech, then the protocol is not simply representative of meaning. It is, rather, *an agent in the production of meaning*" (p. 240). (See also Vygotsky 1997.)

In a sociocultural theory of mind, verbalization is conceived of as a tool that enables ***changes*** in cognition. Speech serves to mediate cognition. Initially an exterior source of physical and mental regulation, speech takes on these regulatory functions for the self. One's own speech (through a process of internalization) comes to regulate, organize, and focus an individual's own mental activities (e.g., Luria 1959, 1973; Sokolov 1972). Clark (1998) refers to this role of language in human cognition as "attention and resource allocation" (p. 172): speech helps us to focus our attention, monitor and control our behaviour.

Another way in which language intersects with the activities of the mind is that it allows ideas to be retained and held up for inspection by the self

and others; it allows ideas to move between people. Such movement allows for "the communal construction of extremely delicate and difficult intellectual trajectories and progressions. . .moreover, the sheer number of intellectual niches available within a linguistically linked community provides a stunning matrix of possible inter-agent trajectories" (Clark 1998: 172).[1] This is what Vygotsky (1978) meant by proposing that the source of learning is social; and what Salomon (1993) and others mean by "distributed cognition".[2]

Vygotsky (1987), Barnes (1992), Wells (1999), and others argue that speech can serve as a means of development by reshaping experience. It serves as a vehicle "through which thinking is articulated, transformed into an artifactual form, and [as such] is then available as a source of further reflection" (Smagorinsky 1998: 172), as an object about which questions can be raised and answers can be explored with others or with the self. Language is data, and with language we are able to manipulate ideas, re-organize them, reshape them, transform them, and construct new ones. "The process of linguistic formulation creates the stable structure to which subsequent thinkings attach" (Clark 1998: 177). As Vygotsky (1987) argued, language is a tool which permits our mind to engage in a variety of new cognitive operations and manipulations. "It enables us, for example, to pick out different elements of complex thoughts and to scrutinise each in turn. It enables us to 'stablise' very abstract ideas in working memory.[3] And it enables us to inspect and criticize our own reasoning in ways that no other representational modality allows" (Clark 1998: 178).

This being so, verbal protocols – which mediate the articulation of cognition – have the power to influence cognition. They exert this influence in three ways. First, the process of verbalization itself transforms thought, drawing attention to some aspects of the environment and not others, solidifying meaning, and creating an observable artifact. Secondly, as an observable artifact, it can be reflected upon, questioned, manipulated and restructured. And thirdly, internalization of this now differently understood externalized artifact may occur. What this implies is that verbal protocols not only potentially transform thinking, focussing it in highly specific ways, but also are the sources of changes in cognition. In other words, speech mediates learning and development.

Some relevant research

In recent research, we have been attempting to demonstrate that speaking mediates second language learning. Our initial work in this area with grade 7 and 8 French immersion students has shown that speaking in the form of dialogue

Table 1. Overview of Swain and Lapkin's (2000) study

Week 1			**Week 2**		
Tuesday	Wednesday	Thursday	Monday	Thursday	Friday
Stage 1 Writing (Pretest)	**Stage 2** Reformulation	**Stage 3** Noticing	**Stage 4** Stimulated Recall	**Stage 5** *Posttests*	**Stage 6** Interviews
Students work together in pairs to write a story in French.	NS of French reformulates story.	Students notice differences between their own story and the reformulated version of their story. This is video-taped.	Video of noticing is shown and students asked to verbalize what they were thinking at the time the video was made.	Students individually rewrite their story.	Students are interviewed individually about the various stages of the study.

mediates second language learning (e.g., Swain & Lapkin 1998). We have called the particular form of dialogue we have been investigating "collaborative dialogue".[4] We have shown that through collaborative dialogue, learners come to know what they do not know or know only partially about language, focus their attention on aspects of language that are problematic for them, raise questions about those problematic aspects of language and respond to those questions (formulate and test hypotheses), and, in so doing, consolidate their existing knowledge or create knowledge that is new for them.

In our more recent work (e.g., Swain & Lapkin 2002, in press), we have added stimulated recalls to our research procedures to try to understand learning processes better from the learners' perspectives. We have also incorporated a pretest/posttest design. This is shown in Table 1. Between the pretest and posttest, students (Grade 7 French immersion students) examine a reformulation of a story they have written and are asked to notice what differences there are between the story they wrote and the reformulated version of their story. While the students are engaged in noticing the differences, they are videotaped. Next, the students see themselves noticing the differences between their own story and its reformulated version, and the tape is stopped each time the students noticed something and asked to tell what they were thinking at the time.

An example of what happens is shown in Table 2. In their story, Nina and Dara, two grade 7 French immersion students, had written "…*elle s'endore sans bruit*" – meaning "she fell asleep without a sound" and the reformulator

changed this to "*. . .elle s'endore dans le silence*" – meaning "she fell asleep in the silence", which altered the meaning of Nina and Dara's original story. Nina and Dara's version puts the emphasis on **how** the girl in their story falls asleep, that is, without a sound; and the reformulator's version highlights the state of the room, which is silent. Nina and Dara noticed this change, and during the stimulated recall, Nina articulates the difference in meaning. She says: "*I think sans bruit is more, she, she fell asleep and she didn't make any noise. But silence is like everything around her is silent.*" Here she puts into words the difference in meaning between their version and the reformulator's version, and when they later individually rewrite their story, although they make use of the reformulator's word, "*silence*", they cleverly manage to preserve their original meaning, Dara by using "*silencieusement*" and Nina by using "*en silence*" – both meaning "silently". Later, when interviewed, Nina makes the more general point about the feedback that they received through the reformulation, "*. . .some of them* [*the reformulations*]*, they seemed like they changed the story sort of and it wasn't really ours.*" – which explains why and how they used this specific aspect of the reformulator's feedback.

The key point here is that by verbalizing the difference in meaning between the two versions, the students were able to accommodate the feedback they received, yet preserve their own meaning – a complex cognitive task. Their final written versions (posttests) would seem to have been affected by both the reformulation itself, and a clear articulation of the differences between their meaning and the one that they felt was being imposed on them. It seems likely that what the students said affected their posttest results. However, our research design did not let us separate out the effect of the feedback from that of the stimulated recall – something that will need to be done to understand the effect of such verbalization.

In a recently completed Ph.D. study, because of the vagaries of doing classroom research in real time, Nabei (2002) separated the effects of the feedback (oral recasts from the teacher) from the stimulated recall. An overview of her study is shown in Table 3. As shown in Table 3, Nabei videotaped the teacher-student interaction in an EFL college class in Japan and took note of all the recasts that the teacher made of student utterances. Based on the specific recasts that occurred, she developed test items that she then administered to the students. After that, she held a stimulated recall session with each student individually in which she showed each student each recast episode and asked them what they were thinking at the time. But, due to time constraints, it was not always possible to show each and every recast episode. This cycle of videoing the class interaction, developing and administering posttest items based on

Table 2. Sans bruit/dans le silence: Nina and Dara

Week 1 Tuesday	Wednesday	Thursday	**Week 2** Monday	Thursday	Friday
Stage 1 Writing-(Pretest)	**Stage 2** Reformulation	**Stage 3** Noticing	**Stage 4** Stimulated Recall	**Stage 5** *Posttests*	**Stage 6** Interviews
Nina and Dara write: *. . . elle s'endore sans bruit.* *	Reformulator writes: *. . . elle se rendort dans le silence.*	D: (reading from the reformulated version) *Se rendort en silence.* N: *Qu'est-ce qu'on a mis?* D: *Sans bruit.* N: *Okay.*	N: I think sans bruit is more, she, she fell asleep and she didn't make any noise. But silence is like everything around her is silent.	Nina writes: *. . . elle se rendore en silence.* Dara writes: *. . . elle s'endore silencieusement.*	N: Some of them (the reformulations), they seemed like they changed the story sort of and it wasn't really ours.

* italics = what students and reformulator wrote

Table 3. Overview of Nabei's (2003) study

cycle repeated six times

Teacher-Student in-class interaction videotaped.	Teacher recasts of student utterances (recast episodes) isolated.	Test items developed based on recast episodes.	First posttest administered.	Stimulated recall session. Students shown recast (and other) episodes. Due to time constraints not all recast episodes were shown.	Second posttest (combination of all first posttests) administered 3 weeks after end of 6 cycle.

the specific recasts provided by the teacher, then conducting stimulated recall sessions was repeated weekly for 6 weeks. Three weeks after completing these 6 cycles of data collection, Nabei gave the students the same test items again – constituting the second posttest – in order to determine if the immediate learning from the recasts was maintained over time. Because some recast episodes were not shown to the students in their stimulated recall sessions, this means that in this second posttest, some of the items were ones where students had spoken about the episode from which the test item was constructed, while for other items, no such stimulated recall took place.

Table 4. Percentage correct items on posttests

	Items with no associated stimulated recall	Items with associated stimulated recall
Posttest 1	68% (19/28)	57% (78/138)
Posttest 2	44% (12/28	64% (88/138)

The results are shown in Table 4. In the first posttest, given prior to any stimulated recall, there are two findings. On the items for which there was never any associated stimulated recall, the average correct score was 68%, whereas on the items on which there was later related stimulus recall data, the average correct score was 57%. This suggests that the items where the learners provided a stimulated recall protocol may have been those that were more difficult for them. The reason I suggest this is that, in the second posttest, on the items where the learners had provided a stimulated recall protocol, their average correct responses went from 57% to 64%; whereas on the items where a stimulated recall had not taken place, the learners' average correct responses went from 68% to 44%. This suggests that the stimulated recalls not only helped the students to maintain over time what they had learned from the recast feedback in class, but also to further develop their knowledge.

Adams (2003) replicated the Swain and Lapkin (2002) study with university students of Spanish using a research design which made it possible for her to separate the effects of task repetition alone (students only wrote the pretest and the posttest), noticing (students, after writing the pretest, compared their writing to that of a reformulated version, then wrote the posttest), and stimulated recall (students wrote the pretest, noticed the differences between their writing and the reformulation of it, and then immediately after the noticing session, the students recalled what they were thinking at the time of their noticing, stimulated by listening to a recording of their noticing session). The posttest score for each learner was calculated as a proportion of reformulations that were incorporated in a more target-like form to the total number of reformulations. Both the Noticing Group and the Noticing + Stimulated Recall Group significantly outperformed the Task Repetition Group. In a further analysis, when the proportion of more target-like incorporated reformulations to reformulations *that the learners had reported noticing* were calculated, the Noticing + Stimulated Recall group significantly outperformed the Noticing Group. These findings suggest that noticing the feedback provided by the reformulation had an effect on the final scores students obtained,

and that the stimulated recall had an impact above and beyond that of noticing the feedback.

What these results show is that speaking, in the form of a stimulated recall, positively affected the performance of language learners. Information processing theory might claim the results are explained by the participants having had more "time on task" and during that time on task, they were given an additional exposure to the information about the correct response and they attended to that information, strengthening their memory traces. For example, Ericsson and Simon (1993) reported on a study on vocabulary learning (Crutcher 1990) where half the items where followed by retrospective reports, and the findings showed that retention was better for those items that called for retrospective reports during learning. "This finding was expected, as the retrospective reports involve an additional retrieval of the memory trace linking the vocabulary pair and hence serve as an additional rehearsal and strengthening of the memory trace" (Ericsson & Simon 1993: xxi).

Sociocultural theory claims the results find *their source* in the verbalization itself. Speaking was not just a report of thought (memory), but it shaped and brought thought into existence.

In other educational domains such as mathematics and science, language has been shown to mediate the learning of conceptual content (e.g., Newman, Griffin, & Cole 1989). The Russian developmental psychologist, Talyzina (1981) demonstrated in her research the critical importance of language in the formation of basic geometrical concepts. Talyzina's research was conducted within the theoretical framework of Gal'perin (1969). With Nikolayeva, Talyzina conducted a series of teaching experiments (reported in Talyzina 1981). The series of experiments dealt with the development of basic geometrical concepts such as straight lines, perpendicular lines, and angles.

Three stages were thought to be important in the transformation of material forms of activity to mental forms of activity:[5] a material (or materialized) action stage; an external speech stage; and a final mental action stage. In the first stage, students are involved in activities with real (material) objects, spatial models or drawings (materialized objects) associated with the concepts being developed. Speech serves primarily as a means of drawing attention to phenomena in the environment (p. 112). In the second stage, speech "becomes an independent embodiment of the entire process, including both the task and the action" (p. 112). This was instructionally operationalized by having students formulate verbally what they carried out in practice (i.e., materially) – a kind of ongoing think-aloud verbalization.[6] And in the final mental action stage, speech is reduced and automated, becoming inaccessible to self-observation

(p. 113). At this stage, students are able to solve geometrical problems without the aid of material (or materialized) objects or externalized speech.

In one of the series of instructional studies conducted by Talyzina and her colleagues, the second stage – the external speech stage – was omitted. The students in the study were average-performing, grade five students in Russia. The performance of students for whom the external speech stage was omitted was compared to that of other students who received instruction related to all three stages. The researchers concluded that the omission of the external speech stage inhibited substantially the transformation of the material activity into a mental one. They suggest this is because verbalization helps the process of abstracting essential properties from nonessential ones, a process that is necessary for an action to be translated into a conceptual form (p. 127). Stated otherwise, verbalization mediates the internalization of external activity.

Holunga (1994) conducted a study concerned with second language learning that has many parallels to those carried out by Talyzina and her colleagues. Holunga's research involved adults who were advanced second language learners of English. The study was set up to investigate the effects of metacognitive strategy training on the oral accuracy of verb forms. The metacognitive strategies taught in her study were predicting, planning, monitoring and evaluating (Brown & Palincsar 1981). What is particularly interesting in the present context is that one group of her learners was instructed, as a means of implementing the strategies, to talk them through as they carried out communicative tasks in pairs. This group was labelled the metacognitive with verbalization, or MV, group. Test results of this MV group were compared to those of a second group which was also taught the same metacognitive strategies, and which carried out the same communicative tasks in pairs. However, the latter group was not instructed to talk about the metacognitive strategies as they implemented them. This group was called the metacognitive without verbalization, or M, group. A third group of students, included as a comparison group (C group), was also provided with language instruction about the same target items, verbs. Their instruction provided opportunities for oral language practice through the same communicative tasks completed by the other students, but the students in this group were not taught metacognitive strategies. Nor were they required to verbalize their problem-solving strategies. Each group of students in Holunga's study received a total of 15 hours of instruction divided into 10 lessons. Each lesson included teacher-led instruction plus communicative tasks to be done in pairs.

The students in this study were tested individually, first by being asked a series of discrete-item questions in an interview-like format, and secondly

by being asked three open-ended questions in which learners would give their opinions, tell a story and imagine a situation. The questions were designed to elicit specific verb forms concerning tense, aspect, conditionals and modals, and were scored for the accuracy of their use. A pretest, posttest and delayed posttest were given. The delayed posttest was administered four weeks after the posttest.

The data were analyzed statistically as four separate tests: The analyses revealed that the MV group made significant gains from pre- to posttests in all four tests; the M group made significant gains in only the discrete-item questions. And the C group showed no improvement on any of the four tests. Furthermore, both the MV and M groups' level of performance at the posttest level was maintained through to the delayed posttests four weeks later. A second set of analyses indicated that both experimental groups performed better than the comparison group on all four tests. Furthermore, the MV group's performance was superior to that of the M group.

In summary, although those students who were taught metacognitive strategies improved the accuracy of their verb use relative to a comparison group that received no such instruction, students who were taught to verbalize those strategies were considerably more successful in using verbs accurately.

Interpreting these findings through the lens of Talyzina's theoretical account suggests that for the MV group, external speech mediated their language learning. Verbalization helped them to become aware of their problems, predict their linguistic needs, set goals for themselves, monitor their own language use, and evaluate their overall success. Their verbalization of strategic behaviour served to guide them through communicative tasks allowing them to focus not only on "saying", but on "what they said". In so doing, relevant content (i.e., the artifact that speech produces) was provided that could be further explored and considered. Test results suggest that their collaborative efforts, mediated by dialogue, supported their internalization of correct grammatical forms. (See also Huang 2004; Negueruela 2003; Swain 2005.)

The studies reviewed above suggest that verbalization (speaking), particularly the verbalization that takes place as one reflects (the "saying") on the artifact created by speech (the "said") plays a significant role in the development and learning that was demonstrated to have taken place.

Implications

What are the implications for research of the different interpretations of verbalizing (speaking) made by information processing and sociocultural theorists? The inferences one anticipates drawing from verbal protocols are not dissimilar across these two theoretical perspectives. Both aim to develop claims about the higher mental processes participants make use of in carrying out a specific task, e.g., solving a mathematical problem, a logical reasoning problem, or a language problem. Information processing researchers use verbal protocols to develop and test "detailed information processing models of cognition, models that can often be formalized in computer programming languages and analyzed by computer simulation" (Ericsson & Simon 1993: 220). Sociocultural theorists also use verbal protocols to discover mental processes underlying task performance[7] (Wertsch 1980; Donato & Lantolf 1990; Swain 2000, 2001). Both research agendas try to explain how and why people think and act: information processing by prediction (projecting into the future based on current behaviour); sociocultural theory by genetic analysis (analysis of the process(es) being formed).

Information processing theorists view verbal reports within the limited constraints of individual task performance, seeking to identify similarities within or across group behavior, whereas sociocultural theories take a broader perspective on such data, attempting to explain them in reference to long-term personal histories. Looking at verbal reports from this broader perspective, temporally and circumstantially, people are interacting and changing with all they say and think, regulating themselves and the world around them. Verbal reports as indications of what people are attending to as they try to complete a short task make people seem static and disembodied from their long-term individual development and their social relations, and focused just on the goals associated with that task. Both theories are concerned with learning, but the extent of the perspective each adopts is different (Cumming, personal communication, November 25th, 2003).

The heart of the matter lies in what is considered to be the relationship between thought and language. For information processing theorists, the two are the same, and verbal protocols are a direct encoding of the heeded thought (Ericsson & Simon 1993). For sociocultural theorists, thought is mediated by the cultural artifacts of our situated being. One of the most important cultural artifacts is language. Through the process of speaking – the articulation and completion of thought – our attention may be refocussed, the boundaries of

thought may be expanded or limited, new ideas may be created, etc. In other words, verbalization changes thought, leading to development and learning.

Returning now to the original question, what do verbal protocols represent? Do they represent cognitive "dumps"? or are they, instead, part of the process of cognitive change, that is, of learning and development? Information processing theory supports the former view. The research I presented in this chapter suggests that the latter view is a strong possibility. It is certainly a matter that needs to be closely studied. If, as I have suggested, speaking and cognitive change can be closely allied, then this needs to be taken account of in any study which makes use of verbal protocols. Verbal protocols cannot be used neutrally as a method of collecting data, but instead they need to be considered as part of the "treatment" when making claims about learning and development. Research tools such as think alouds and stimulated recalls should be understood as part of the learning process, not just as a medium of data collection (Smagorinsky 1998; Swain 2005). Think alouds and stimulated recalls are not, as some would have it, "brain dumps"; rather they are a process of comprehending and reshaping experience – they are part of what constitutes development and learning.

Notes

* I would like to thank Micheline Chaloub-Deville, Louis Chen, Alister Cumming, David Ishii, Penny Kinnear, Jim Lantolf, Toshiyo Nabei, Sharon Lapkin and Harry Swain for reading earlier draft(s) of this chapter. Their comments led me to infer more and generalize less.

1. Clark is essentially a connectionist. In this regard, it is interesting how closely many of his claims (1998) echo the words of Vygotsky.

2. Even in the case of a person acting alone, cognition is still distributed because of internalization. This is not a point forcefully made in Salomon (1993). (Personal communication, Lantolf, December, 2003.)

3. Clark (1998) suggests that "for certain very abstract concepts, the *only* route to successful learning may go via the provision of linguistic glosses. Concepts such as charity, extortion and black hole seem pitched too far from perceptual facts to be learnable without exposure to linguistically formulated theories. Language may thus enable us to comprehend equivalence classes that would otherwise lie forever outside our intellectual horizons" (p. 170).

4. Collaborative dialogue is dialogue in which speakers are engaged in problem-solving and knowledge-building – in this case, solving linguistic problems and building knowledge about language.

5. I.e., internalization – conversion of objective to idealized activity (Gal'perin 1969).

6. The difference is that in a think aloud the individual is asked to say what they personally are doing, but in Talyzina the individual's speech is supposed to reflect the conceptual understanding of the process provided by the instructor and materialized in some form or other.

7. What is going on in speaking is a genetic process and so we should be able to improve this process by modifying the things that people say about what it is they are doing (Gal'perin 1967). Gal'perin argues that through conceptualization, materializaton, verbalization and internalization we can hasten development.

References

Adams, R. (2003). L2 output, reformulation and noticing: Implications for IL development. *Language Teaching Research, 7,* 347–376.

Barnes, D. (1992). *From communication to curriculum* (2nd ed.). Portsmouth, NH: Heinemann.

Brown, A. & Palincsar, A. (1981). Inducing strategic learning from texts by menas of informed, self-controlled training. *Topics in learning and learning disabilities, 2,* 1–17.

Carruthers, P. & Boucher, J. (Eds.). (1998a). *Language and thought: Interdisciplinary themes.* Cambridge: CUP.

Carruthers, P. & Boucher, J. (1998b). Introduction: Opening up options. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 1–18). Cambridge: CUP.

Chaloub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 2,* 369–383.

Clark, A. (1998). Magic words: How language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 162–183). Cambridge: CUP.

Crutcher, R. J. (1990). *The role of mediation in knowledge acquisition and retention: Learning foreign vocabulary using the keyword method.* Tech Report No. 90–10. Boulder: University of Colorado, Institute of Cognitive Science.

Cumming, A. (1990). Metalinguistic and ideational thinking in second language composing. *Written Communication, 7,* 482–511.

Donato, R. & Lantolf, J. P. (1990). Dialogic origins of L2 monitoring. *Pragmatics and Language Learning, 1,* 83–97.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). Cambridge, MA: The MIT Press.

Ericsson, K. A. & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5,* 178–186.

Gal'perin, P. Ya. (1967). On the notion of internalization. *Soviet Psychology*, *5,* 28–33.

Gal'perin, P. Ya. (1969). Stages in the development of mental acts. In M. Cole & I. Maltzman (Eds.), *A handbook of contemporary Soviet psychology.* New York, NY: Basic Books.

Gass, S. M. & Mackey, A. (2000). *Stimulated recall methodology in second language research.* Mahwah, NJ: Lawrence Erlbaum.

Holunga, S. (1994). The effect of metacognitive strategy training with verbalization on the oral accuracy of adult second language learners. PhD Dissertation, University of Toronto (OISE), Toronto.

Huang, L. (2004). A little bit goes a long way: The effects of raising awareness of strategy use on advanced adult second language learners' strategy use and oral production. PhD Dissertation, University of Toronto (OISE), Toronto.

Lantolf, J. (Ed.). (2000). *Sociocultural theory and second language learning.* Oxford: OUP.

Lantolf, J. (2003). Intrapersonal communication and internalization in the second language classroom. In A. Kozulin, V. Ageev, S. Miller, & B. Grindis (Eds.), *Vygotsky's educational theory in cultural context* (pp. 349–347). Cambridge: CUP.

Lantolf, J. & Appel, G. (Eds.). (1994). *Vygotskian approaches to second language research.* Norwood, NJ: Ablex.

Luria, A. R. (1959). The directive function of speech in development and dissolution. *Word, 15,* 341–352.

Luria, A. R. (1973). *The working brain.* New York, NY: Basic Books.

Mackey, A. (2002). Beyond production: Learners' perceptions about interactional processes. *International Journal of Educational Research, 37,* 379–394.

Nabei, T. (2002). Recasts in classroom interaction: A teacher's intention, learners' awareness, and second language learning. PhD Dissertation, University of Toronto (OISE), Toronto.

Negueruela, E. (2003). A sociocultural approach to teaching and researching second languages: Systemic-theoretical instruction and second language development. PhD Dissertation, Pennsylvania State University, State College, PA.

Newman, D., Griffin, P., & Cole, M. (1989). *The construction zone: Working for cognitive change in school.* Cambridge: CUP.

Salomon, G. (Ed.). (1993). *Distributed cognitions: Psychological and educational considerations.* Cambridge: CUP.

Searle, J. R. (2002). *Consciousness and language.* Cambridge: CUP.

Smagorinsky, P. (1998). Thinking and speech and protocol analysis. *Mind, Culture, and Activity, 5,* 157–177.

Smagorinsky, P. (2001). Rethinking protocol analysis from a cultural perspective. *Annual Review of Applied Linguistics, 21,* 233–245.

Sokolov, A. (1972). *Inner speech and thought.* New York, NY: Plenum Press.

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 97–114). Oxford: OUP.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*, 319–346.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471–484). Mahwah, NJ: Lawrence Erlbaum.

Swain, M. & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *Modern Language Journal, 82,* 320–337.

Swain, M. & Lapkin, S. (2002). Talking it through: Two French immersion learners' response to reformulation. *International Journal of Educational Research, 37,* 285–304.

Swain, M. & Lapkin, S. (in press). "Oh, I get it now!" From production to comprehension in second language learning. In D. M. Brinton & O. Kagan (Eds.), *Heritage language acquisition: A new field emerging.* Mahwah, NJ: Lawrence Erlbaum.

Talyzina, N. (1981). *The psychology of learning.* Moscow: Progress Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.)). Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1986). *Thought and language.* Cambridge, MA: The MIT Press.

Vygotsky, L. S. (1987). *The collected works of L. S. Vygotsky.* Volume 1. *Thinking and speaking.* New York, NY: Plenum Press.

Vygotsky, L. S. (1997). *The collected works of L. S. Vygotsky,* Volume 4. *The history of the development of higher mental functions.* New York, NY: Plenum Press.

Wells, G. (1999). *Dialogic inquiry: Toward a sociocultural practice and theory of education.* Cambridge: CUP.

Wertsch, J. V. (1980). The significance of dialogue in Vygotsky's account of social, egocentric, and inner speech. *Contemporary Educational Psychology, 5,* 150–162.

Wertsch, J. V. (1985). *Vygotsky and the social formation of mind.* Cambridge, MA: Harvard University Press.

Wertsch, J. (1991). *Voices of the mind: A sociocultural approach to mediated action.* Cambridge, MA: Harvard University Press.

Wertsch, J. & Tulviste, P. (1996). L. S. Vygotsky and contemporary developmental psychology. In H. Daniels (Ed.), *An introduction to Vygotsky* (pp. 53–74). London: Routledge.

Functional grammar

On the value and limitations of dependability, inference, and generalizibility

Diane Larsen-Freeman
University of Michigan

In this paper, I address the themes of this volume from a functional grammar perspective. I explain the importance of being able to make inferences about the form, meaning, and use of grammar structures that rest on dependable data. Multiple data sources increase the chances that the data are dependable, i.e., consistent from one context to the next.

However, even by using complementary data sources, it is still the case that inferences can only be partial and provisional. In addition, it is impossible to define an optimal level of generalizability for our inferences apart from the purpose to which the generalization is to be put and for the particular audience that it is intended. Then, too, for certain applied linguistic purposes, optimal levels of generalizability cannot be foreordained, but, instead, will have to be negotiated.

Introduction

I bring a functional grammar perspective to bear on the way people mobilize their grammatical resources in order to make meaning, to maintain the flow of information, to manage interpersonal relationships, and to position themselves socio-politically, among other things. While the findings of any linguistic investigation must always be provisional and partial, for reasons I discuss below, it is important nevertheless to attempt to make claims generalizable, within limits. Certainly I would want any grammatical explanation to apply beyond the specific data from which it was inferred. If claims were limited to a single data set, they would not be very useful in accomplishing the purposes to which I, as an applied linguist, want to put them, which are:

- to inform the identification of the language acquisition/learning challenge of language students (what is the nature and challenge of that which is being learned/acquired?)
- to better understand the various processes contributing to, or interfering with, meeting the learning/acquisition challenge
- to adopt pedagogical strategies, to design materials, and to educate teachers, all informed by an understanding of the learning challenge and learning processes.

I wrote "inform the identification of the learning/acquisition challenge," not simply "identify," because the nature of the challenge, the processes, and the strategies depend as much on who the learners are as on what it is that they are learning. In other words, the learning challenge is not purely a linguistic one. For this reason, I will illustrate how the two foci – the "what" of the object of learning and the "who" of the learner – come together later on in this chapter.

In order to study the "what" or functional grammar, my students and I have used contextual analysis (Celce-Murcia 1980), a research methodology that employs complementary procedures to investigate the patterned use of grammatical structures. In this chapter, I first discuss the object of study – what I seek to construct knowledge about – then I turn to the methodology that is used to create the knowledge. Following this, I expand upon my assertion above – that any claims the methodology yields must be provisional and partial. Finally, I consider the difficult issue of what the optimal level of generalizability for applied linguistics research ought to be.

The object and purpose of the research

The research I describe in this chapter seeks to address three questions:

- How is a particular linguistic structure formed, i.e., its morphosyntax and its phonology, when the latter is grammatically relevant? The answer to this question not only includes paradigmatic grammatical relations, but also the syntagmatic relationship to items that precede or follow it in the discourse.
- What does it mean, i.e., its semantics? Another way to pose this question is to ask what the "essential" or referential meaning of the structure is, beyond its use in a particular context.
- When/why is it used, i.e., what are the salient pragmatic factors governing its use in particular contexts? Why is this form used and not another form

that has a similar meaning? To answer this question, one must examine the role that the structure under investigation plays in the overall discourse. For example:

> Does it initiate, terminate, or continue episodes?
> Does it contribute to information flow, thematic structure, discourse coherence or cohesion?
> Does it signal affective or represent socio-interactional factors that are present in the context?
> Etc.

In applied linguistics, often a binary distinction is made between form and function or between form and meaning, rather than the ternary one I have just proposed. I think that binary distinctions overlook an important dimension to learning and using grammar, which I try to account for with a tripartite framework. For example, students of English can learn that *-ed* is used as a past tense form to mark a completed action, but knowing both form and meaning will be insufficient in situations of use where speakers are forced minimally into choosing between using the present perfect or the past for this same meaning,

> I have recently returned from Cyprus.
> I recently returned from Cyprus.

which, in American English, are both permissible sentences. The choice here presumably depends on one's orientation to the event (see, for example, Larsen-Freeman, Kuehn, & Haccius 2002). My approach thus rests on a fundamental linguistic principle: the linguistic system is holistic, a choice exists among linguistic forms, and when one form is chosen over another, there are always semantic or pragmatic consequences.

To illustrate the framework in which this research is conducted, consider the case of the existential *there* in English.

How is it formed?

Existential *there* is a single, invariant, free morpheme, which occupies the subject position in an English sentence. The logical subject, the one that governs the verb, follows the verb, which is frequently a form of *be*. Following the logical subject is often, though not always, a locative adverbial. Existential *there* is unstressed, and therefore phonologically reduced, which is one of the way to distinguish it from the pro-adverbial *there*.

What does it mean?
It is called the existential *there* because it establishes a mental space in which some entity exists or is to be located (Celce-Murcia & Larsen-Freeman 1999). It, therefore, has a different (although, of course, related – see Lakoff 1987) meaning from the pro-adverbial *there*, which is used deictically with physical space.

When or why is it used?
It has a presentational function; it brings an element into awareness (Langacker 1991). By filling the subject slot with *there*, a speaker can put an entire proposition in the end-focus position in a sentence, a spot reserved for new information. In this way, its function is to establish or to maintain the thematic focus.

This three-dimensional analysis is the product of an inferential process linguists and applied linguists undertake based on the language performance of speakers of English. For applied linguistics purposes, as I wrote earlier, not only the "what," but also the "who" must be considered in order to inform the identification of the learning challenge. One implication of this assertion is that a parallel inferential process should be conducted with language learner data. Inferences based on the language performance of both native speakers and learners are always subject to review, of course, either on their own terms or as additional data are gathered.

A brief, selective treatment of the literature on English language learner performance, taking into account only one learner factor – the native language of learners – suggests that speakers of topic-comment languages, Japanese, for example, transfer the topic-comment structure of their native language to their English interlanguage, thus avoiding the use of *there* where speakers of English would use it. In other words, instead of saying

> There are 27 students in Taro's school.

They say

> *Taro's school is 27 students.
> *Taro's school students are 27.
> *In Taro's school students are 27.
>
> (Examples from Sasaki 1990, cited in Celce-Murcia & Larsen-Freeman 1999)

Japanese speakers of English also make use of the English verb *have*, which allows them to preserve the topic-comment word order of Japanese. While the

form that results from the application of this strategy is not inaccurate, it could be pragmatically inappropriate, depending on the context.

?Taro's school has 27 students.

Another problem for speakers of topic-comment languages is the formation of "pseudo-relatives" (Yip 1995). For example, Chinese students use the existential *there* in a way that makes it appear that they have difficulty with English relative clauses (Schachter & Rutherford 1979: 3). They say

*There were a lot of events happen in my country.

However, Rutherford (1983) infers that such ungrammatical learner utterances do not stem from failure to use a relative pronoun or a participle form, but rather can follow learners' incorrectly perceiving there to be a functional equivalence between *there* and a topic introducer in Mandarin Chinese.

Finally, from the questions they ask, we know that most ESL/EFL students are perplexed by the fact that two different sentence constructions are possible in English, which appear to have the same meaning (although, of course, they have entirely different uses).

There are 27 students in Taro's school.
27 students are in Taro's school.

Inferring from the students' linguistic behavior and their questions, and invoking the challenge principle (Larsen-Freeman 1991, 2003), we are left to conclude that for these students, it is the use of existential *there* that affords the greatest long-term learning challenge. Of course, students need to learn its form and its meaning as well, but since it is impossible to teach everything about a given structure, let alone to teach it at one point in time, our conclusion tells us where to direct students' attention in an attempt to be maximally effective with the limited instructional time we have available. It will also inform the choice of pedagogical strategies. For example, our analysis suggests that showing students a picture and asking them to make sentences about what they see using *there* is not an effective strategy for addressing the learning challenge most students will confront, which is to learn to use *there* to introduce new information. This is because if the students and teacher are looking at the same picture, students would be using *there* to make sentences about known information, and such a teaching strategy would therefore mislead students about the use of *there.*

In terms of teacher education, teacher learners need to be able to answer the three questions about the form, the meaning, and the use of *there.* Further-

more, in order to take advantage of learners' cognitive abilities, teachers might think in terms of reasons, not rules (Larsen-Freeman 2000). For example, the well-known rule that states that the logical subject, which follows the verb in sentences with existential *there*, is indefinite can easily be explained not as an arbitrary rule of form, but as a reason: the logical subject is indefinite because it conveys new information.

I have gone to some length to illustrate what it is that I do and the purposes to which the results are put. It is time now to be explicit about the general themes of this volume. For the research program that I am describing in this chapter, generalizability is desirable. Researchers using this approach seek to be able to make inferences, such as the function of X is Y, or X collocates with Y, or X means Y – based on performance data that are dependable, i.e., that are consistent from one instance to the next.

A research methodology for functional grammar

Contextual analysis is a research methodology that is used to examine written and spoken data in an attempt to account for the form, the meaning, and the use of linguistic forms in texts. To my knowledge, the research has only examined native English speaker texts, but there is nothing in the methodology itself that would prohibit its being used to analyze other varieties of English, English as a lingua franca, or for other languages. Indeed, given that many learners do not aspire to native English speaker performance, it would no doubt be desirable to have it apply to whichever language or language variety is to be learned.

There are five steps to the methodology (based on Celce-Murcia 1980, 1990):

1. Choose a linguistic structure to examine. One's choice may be informed by observation of usage data, or by being unhappy with simplistic pedagogical "rules of thumb," or by seeing one's language students struggle with a particular structure, etc.
2. Review the literature. Read published linguistic and pedagogical accounts that others have written about the structure chosen.
3. Make use of intuitions. Intuitional data might consist of grammaticality judgments (if form is the issue) or of contrasting two made-up sentences or texts if meaning or use is the issue (cf. *She isn't much fun. She isn't a lot of fun*).

4. Observe the structure as it is used in natural written and spoken discourse of native speakers. Try to determine why a particular structure is used in a particular context and why it is used rather than some other structure with a similar meaning.
5. Test hypotheses generated from steps 2–4 by eliciting judgments or data samples from other speakers.

It can be seen, then, that because there are three primary, complementary data sources, intuitional, observational, and elicited, contextual analysis seeks to maximize the dependability of linguistic data from which patterns can be observed and for which generalizable explanations can then be inferred. I would like to turn next to examining the three data sources and what each contributes to this goal.

Intuitional data

Using speaker intuitions about constructed examples is a time-honored practice in linguistics research. In fact, since the rejection of behaviorism, introspecting about linguistic structures has been a most favored heuristic of linguists (Gass & Mackey 2000: 10). Indeed Gass and Mackey quote Bard, Robertson, and Sorace (1996: 32) as stating that "For many linguists, intuitions about the grammaticality of sentences comprise the primary source of evidence for or against their hypotheses."

However, intuitions can be unreliable or undependable when looking for typical patterns because humans tend to notice the unusual more than the typical (Biber, Conrad, & Reppen 1998), and because inferences based on one linguist's intuitions may not be very generalizable for the purposes for which we would like to use our findings.

Further, although

> . . .speaker intuitions about constructed examples are an invaluable tool, their use requires at least the following: an acceptance and appreciation of the cline of acceptability and interspeaker variability that is typically associated with such examples; an understanding of the nature of "deviance" from linguistic norms; and, most generally, some serious reflection on what such judgments actually tell us. But even with such judicious use, intuitions about constructed data cannot be treated as the sole, or even primary, source of evidence as to the nature and properties of the linguistic system.
>
> (Barlow & Kemmer 2000: XV)

So although contextual analysis makes use of intuitional data, for the sake of dependability and generalizability, it complements them with data from two other sources.

Observational data

The source of the observational usage data varies depending on the structure being investigated. It would be appropriate, for instance, to examine conversational transcripts if one were interested in investigating English tag questions. On the other hand, if one wanted to look at the use of the passive voice in scientific writing, then a different source of data would clearly be warranted (Celce-Murcia 1980).

However, modality and genre of usage data are only two factors that are known to affect linguistic choice. Many other factors have been found to influence the meaning or use of linguistic forms, such as the setting, the relationship of speaker and listener or reader and writer, their characteristics (gender, age, educational level), the register of the discourse, whether speech is planned or unplanned, monologic or dialogic, etc. Ideally, for maximum dependability these factors should be systematically addressed as well.

This leads us to a critical problem when relying on observational data. Until recently, it was very difficult to collect sufficient examples to permit systematic sampling of factors that might affect the form, the meaning or the use of particular structures. As it was, researchers would pore over published transcripts of oral data or various types of written data searching for, and hand-counting, instances of the target structure. Given the tedious and time-consuming nature of the observational data collection process, it was clearly not realistic to expect that many of the factors known to affect linguistic choice could be systematically addressed. This limitation compromised the dependability of the data and limited the generalizability of any claims based on them.

More recently, the availability of large computer databases and tools has provided an alternative to such tedious procedures and non-representative data. These resources significantly facilitate the study of the patterned ways in which speakers use the grammatical resources of a language.

To briefly illustrate this point requires returning to the discussion concerning the existential *there*. In his search of a corpus of 450,000 words of spoken text and 300,000 words of written texts of modern, educated, British English speaker usage, Breivik (1981) was able to find very few instances of the type of sentence I gave earlier, where the logical subject with an indefinite determiner occupies subject position (*27 students are in Taro's school.*). One example he did

find, *A headphone bar is also on the first floor*, was in a text describing what one would find on various floors of an electronic equipment store.

> The first floor houses the real heart of the store – the hi-fi departments – but there is much else here also. . .A headphone bar is also on the first floor.

This example is helpful for it contains both types of sentence we are interested in – sentences with and without the existential *there* used for a presentational function. From Brevik's findings, exemplified in this text, we might infer that the paucity of sentences such as the final one in this excerpt can be attributed to the fact that they require that a scene already be established or presupposed to which they then contribute details. This requirement contrasts with the sentence that precedes it, which contains the existential *there*, which has no such requirement, and which here presumably asserts the existence of "much else," thus setting the scene for the sentence without *there*. The frequency distribution of these two sentences in the corpus would be helpful information for our applied linguistic purposes as they would in providing support for inferring a presentational function for existential *there*.

Such inferences must always be provisional, however, subject to refutation as other data are considered, especially when the data include texts from more diverse speakers. While representativeness of a corpus is critical if we are seeking a comprehensive description of a structure, we also have to accept that any absolute comprehensiveness is impossible to achieve. Thus, inductive research, of the sort I have been describing, always seeks to corroborate, expand upon, or to refute that which has preceded it.

To illustrate the tenuous nature of linguistic explanations, consider the work of Kim (1995). Kim was interested in speakers' motivation for using WH-clefts in English. Traditional accounts held that WH-clefts exist to put special focus on new information in the predicate. For example,

> What that amounts to is that they don't keep comparable books.

However, from Kim's examination of spoken data, he was led to conclude that the traditional explanation was incomplete. Only three examples in his data supported this function uniquely. Many of the remaining 73 examples highlighted a speaker-oriented interactional function of WH-clefts, such as a speaker's marking a topic shift for a listener

> What you are saying reminds me of. . .

While in this example the WH-cleft still contrasts old and new information, the following example, drawn from Kim's data, does not highlight the contrast between the types of information so much as it displays speaker affect

What I love is when they're talking about something. . .

Kim's investigation of data from face-to-face and telephone conversations, augmented by talk radio broadcasts and group therapy sessions, was important for expanding the inventory of uses for WH-clefts. However, Kim's data were limited to 73 tokens of such clefts. By contrast, computerized corpora provide access to many more instances of a target structure. For example, a preliminary search of MICASE (Michigan Corpus of Spoken Academic Discourse), a 1.8 million-word database, developed, maintained, and made freely available to the research community by the English Language Institute of the University of Michigan, produced a list of 1052 potential WH-clefts. Admittedly this list contains some interlopers masquerading as clefts, which would have to be eliminated through a refined search (Larsen-Freeman 2004). Nevertheless, even a cursory analysis points to an additional function of WH-clefts in spoken academic discourse, that of prospectively or retrospectively framing instruction (John Swales, personal communication)

What we want to do today is. . .

a function not identified in previous research.

This observation itself points to a potential complication in the use of corpora to advance our research agenda. The larger number of instances of a target structure, which corpora give access to, increases the dependability or stability of observed patterns, but may also reveal a multiplicity of "lower level" findings. Each new corpus, having its own special features, may illustrate more specialized functions for a given structure or more variation in the structure itself. Of course, this is no reason not to continue to look for new functions or structural variation – just a caution that even if were possible to account comprehensively for all the forms, meanings, and uses of a given structure, generating such an exhaustive list would not necessarily meet the needs of applied linguists. If, as applied linguists, we want findings that will usefully inform general, rather than narrowly focused specific purpose, SLA, pedagogy, and teacher education, a long list of forms, meanings, and functions for any one target structure would only be a starting point. The list would still have to be winnowed and shaped in order to manage its contribution to language learning (Widdowson 2000).

Of course, an advantage to starting with a large number of tokens of the target structure, which computer corpora of a certain size make available, is that statistical tests can be used. Such tests can help us to assess the likelihood of whether or not the observed difference or relationship could be due to chance (e.g., using the chi-square test) and then secondly to measure the strength of the relationship between the two variables (e.g, using an association coefficient to say whether or not there is a strong association, such as a collocation, between the two forms). To cite an example (I thank N. Ellis for bringing it to my attention), a chi-square test can provide evidence that *charges* occurs after *bring* more often than randomly, but it cannot tell us how strongly related the two words are even though the result may be statistically significant. By then testing the strength of the association using the uncertainty coefficient, one could find an estimate of the relative reduction of uncertainty for predicting that given the word *bring*, the word *charges* will follow. With an uncertainty coefficient close to 1, there is an indication that a strong association (collocation) has been found and generalizability claims are possible (G. Demetriou on corpora-requests@lists.uib.no). Alternatively, one can calculate an MI (mutual information) score, which, if greater than 2, shows a substantial association between the 2 words (Kennedy 2003). Given a large enough data set, we could also consider the influence of multiple factors at the same time. For example, we could contrast the influence of the matrix verb (which Kim found noteworthy) with the type of speech event.

Thus, linguistic corpora can be enormously helpful in conducting contextual analyses by providing abundant examples of the target structure, thereby increasing the dependability or stability of inferred patterns across language samples, and, in tandem with statistics, may permit greater generalizability. However, at this time, concordancing programs are limited to "low-level" searches of the sort that I have just illustrated with *bring* and *charges*. It is not possible to investigate complex grammatical constructions unless you have programming skills (Biber, Conrad, & Redden 1998: 15) or access to an appropriately tagged corpus. Furthermore, corpus searches simply yield numbers of instances. Interpretation of the instances in order to identify patterns in the data is still necessary. Numbers do not guarantee insight. After all, sometimes a skillful microanalysis of a single text may yield significant insights into the form, the meaning, and the use of the target structure (see Markee chapter in this volume).

To underscore this point, let me discuss some work by Biber, Conrad and Reppen (1998), who investigated "nearly equivalent" *that*-clauses and *to*-clauses from a lexico-grammatical perspective using approximately 4 million

words of academic prose from the Longman-Lancaster Corpus and approximately 5 million words of conversation from the British National Corpus. Using computer programs to count which verbs occurred with each complement type, the researchers show that completely different sets of verbs most commonly control *that*-clauses versus *to*-clauses. *That*-clauses are linked with three matrix verbs – *think*, *say* and *know*, although *believe, mean*, and *tell* also commonly take *that*-clause complements. According to the researchers, their typical use is in reporting what was thought felt, or said. In contrast, the researchers found that the most common verbs occurring with *to*-clauses come from a wider semantic domain. Two verbs are particularly common with *to*-clauses expressing desire (*want* and *like*), while a third is used to express effort (*try*).

While these observations (and others that they go on to make) are helpful, they were not unknown before the use of computerized corpora. Years ago, Givón (1980) discussed the correlation between the matrix verb and the complement it takes. He observed that the main verbs (such as *assume*, *imagine*, *know*, *understand*, *say*) that take ordinary tensed *that*-clause complements mostly denote mental states or attitudes regarding the truth of the proposition in the complement clause. Thus, if someone were to say

> I think that it will rain today.

the speaker is saying something about the nature of his/her belief, and its relative strength. Givón puts such verbs into the cognition-utterance category (Williams in Celce-Murcia & Larsen-Freeman 1999). Furthermore, Bolinger (1968) pointed out still earlier that verbs taking infinitive complements encode future unfulfilled projections, which it would seem the matrix verbs *want, like*, and *try* that Biber et al. point to do. Thus, while corpus linguistics offers a means for testing and well as for discovering linguistic principles, some insightful inferences have long endured and will likely continue to do so.

Another limitation of observational data, whether part of a computer corpus or not, is that they are attested data. Such data show how the linguistic code has been performed at some point in time; they do not show the potential of the code (Widdowson 1990). They also do not show what cannot occur, or what is aberrant if it does occur (Howard Williams, personal communication). For after all, spontaneously produced utterances provide only part of the picture (Gass & Mackey 2000). If one wants to obtain information about why speakers use grammar in the way that they do, it is essential to determine what learners think is possible in the language and what is not – to understand the limits of the system. All that is directly observable is what a learner produces

in writing or speech. Such data do not show what is underlying the observable linguistic behavior. To get at this, elicited data are used.

Elicited data

It could be said that the procedures that I have described in this chapter so far fall into the hypothesis-formation category. Typically, by surveying the literature, consulting one's intuitions, and examining observational data consisting of oral and/or written discourse, researchers are able to come to some tentative answers to the questions posed. However, these still must be subject to tests designed to tap speaker intuitions.

Eliciting data permits one to narrow one's hypothesis space by carefully manipulating the factors that one wants to test. For example, Williams (1996) tested the hypothesis that *nevertheless* is more restricted in its use than *however*. While *however* can almost be used generically wherever attention is drawn to difference, *nevertheless* requires a situation where one is led to expect one thing but finds something different to be true (Celce-Murcia & Larsen-Freeman 1999). In order to test this inference, Williams designed a questionnaire that asked 30 native speakers of English to indicate their acceptance of *nevertheless* or *however* in a set of sentences he constructed. From their responses, Williams found strong acceptance for *nevertheless* only for certain sentences, such as the first one in the following pair:

> John has always been a top math student.
> _____ he failed calculus this quarter. (83 percent acceptance)
> John has always been a top math student.
> _____ he failed history this quarter. (20 percent acceptance)

What this suggests is that *nevertheless* is more restricted in its use than *however*, an observation supported by the fact that within the MICASE corpus, there were 214 instances of *however* in spoken academic discourse, but only 13 instances of *nevertheless.* More in-depth analysis of the actual instances would have to be undertaken, of course, to see if the distinction that is reported above holds. Then, too, the elicitation instrument itself should be validated and its results subject to statistical tests to ensure that any inferences that would be drawn are supported.

On the provisionality and partiality of linguistic inferences

Even by using three complementary data sources, by assuring the validity of the instruments, and by subjecting the results that they yield to the proper statistical tests, it is still the case that any linguistic inference must be provisional for the reason stated earlier: Any inference achieved through an inductive process is always subject to refinement by a counterexample. In addition to being provisional, any linguistic inference is also partial. There are several reasons for this. First of all, language usage is not homogeneous. What will be elicited on a given occasion with a given instrument varies with co-text and context, and it is impossible to be comprehensive in accounting for all the possible permutations of these. Moreover, even the largest possible database is selective. Not every instance of language use has been recorded and is computer-searchable. Then, too, we cannot with any assurance draw conclusions about the grammar of an individual from usage facts about communities (Newmeyer 2003). Modal tendencies can be revealed, but the decision to conform or to innovate is always left up to the individual. Finally, language is constantly on the move, constantly changing, and that part which is pre-systematic is likely to be overlooked.

To expand upon this last point, it could be said that all extant methods, contextual analysis included, give us only a record of attested usage at one point in time. They cannot tell us what transpired in the language up until this point, nor where it is destined. While this may seem obvious and forgivable as the system exhibits stability as well as mutability (Givón 1999), we often underestimate the latter. We therefore miss the perpetually changing, perpetually dynamic nature of language (Larsen-Freeman 1997). As I have argued elsewhere, applied linguistics would no doubt be well-served by thinking of language and its learning in terms of dynamic systems, abundant examples of which occur in the natural world (Larsen-Freeman 2003). When we conceive of language as a static system, we contribute to "the inert knowledge problem" (Whitehead 1929 in Larsen-Freeman 2003). We teach language (grammar) as if it were a static system, and so it becomes one for language learners, who find that they cannot apply what they have learned to novel situations, which is surely the ultimate test of whether something has been usefully acquired.

Optimal level of generalizability

In this chapter, I have staked out the position that generalizability is desirable – that what we would like is to have claims that apply beyond particular in-

stances. While this is true enough, it is important to recognize that there are limits to generalizability as well, at least for applied purposes. What we should be seeking is an optimal level of generalizability, a level admittedly difficult to define apart from the purpose for which it is to be put. For example, linguists have pointed out that the *-ed* morpheme occurs in a number of environments; it is not simply a past tense marker. It occurs, for instance, as a past participle, thus in perfective aspect and passive voice as well, in backshifted verbs in the complement clauses of reported speech, as participial adjectives, in imaginative conditionals, etc. Inferring a meaning underlying all these instances, Knowles (1979) suggests that the *-ed* signals remoteness. While this is an enlightening linguistic observation, a generalization this broad may not be especially helpful for learners of English or only so for learners who have achieved a certain level of proficiency.

The price for increased generalization is increased abstraction. While some linguists pursue the broadest possible generalizations, which are therefore necessarily abstract (for example, Chomsky's 1995 efforts to try to account for the structure of all the world's language with a minimal number of principles), it is difficult to see how such abstract principles are useful for all the purposes to which applied linguists wish to put them.

Conversely, of course, linguistic principles inferred from speakers' performance can be undergeneralized. With so many factors affecting linguistic performance, one may have to resort to statements such as the following: "Structure X is used by female Standard English speakers (ages 45–50) in their conclusions to commencement addresses delivered at major research universities in North America during the years 2000–2005." While the need for such particularizability may be exaggerated, the point is that just as generalizations can be too broad for applied linguistic purposes, they can also be too narrow. Where parsimony (few broad generalizations) or alternatively, comprehensiveness (many narrow generalizations) may be desirable from a theoretical perspective, from an applied perspective, it is hard to see the value of either extreme. A definition of the optimal level of generalizability of linguistic inference is difficult to define with any precision. What we are left with is "not too much so as to become so abstract, not too little so as to put an unnecessary learning burden on the student."

Perhaps, we need to take a lesson from Marr (1982). Marr has pointed out for vision that the same object may be represented at various levels of detail. As a consequence of this, one cannot simply talk about a perceived object possessing some property. Instead, one must talk about whether, given a certain level of detail, it is seen to have this property. By the same token, from an applied

linguistics perspective, one must always inquire as to which purpose an explanation is to be put in order to determine at which level of detail an appropriate level for generalizability is to be aimed.

Of course, the optimal level can never be entirely foreordained. It is in the nature of teaching and learning that the optimal level potentially arises out of negotiations between the teacher and students in a way that past understanding is taken into account and built upon. Still, for materials developers and others who must prepare decontextualized explanations for explicit teaching purposes, questions about the optimal level of generality persist.

To make the point another way, and to introduce another concern, this one having to do with the unit of analysis, let me digress for a moment to consider that fact that an innovation of computer corpora is that they can be searched to reveal the patterns that can be associated with particular lexical items. An example from Hunston and Francis (2000) involves the noun *matter*. It turns out that *matter* is often preceded by an indefinite article and followed by the preposition *of* and a gerund beginning with *-ing*, for example, *a matter of developing skills, a matter of learning a body of information, a matter of becoming able to…*". There is, therefore, little point in treating *matter* as a single lexical item that can be slotted into a general grammar of English. Rather, the word *matter* comes with attendant phraseology.

While this is of interest, particularly if one is pursuing a lexical approach to language acquisition and language pedagogy, it is hard to imagine constructing a coherent research agenda or a comprehensive syllabus based on single lexical items and phrases unless restricted somehow to high frequency lexicogrammatical units or units that are broader in scope. Therefore, in this paper, I have appealed to traditional linguistic categories, verb tense/aspect, tag questions, passive voice, existential *there,* and WH-clefts. While these are broader and commonly recognized, therefore units of convenience, they are not necessarily psychologically real for learners. It is not clear that language students segment the target language in the same way. If not, then whatever can be inferred from native speaker usage data might not be particularly illuminating when analyzing interlanguage data. As has been pointed out before, it may well be that learners' emic perceptions of the target depart radically from any etic linguistic categories.

Conclusion

I have argued in this chapter that researchers in functional grammar should work to ensure that their inferences are drawn from dependable data. One way to do this is to utilize complementary data sources. However, because inferences based on the data are arrived at inductively, they are always subject to counterexamples. Further, it is highly likely that counterexamples will arise in language performance data since the data that are examined are always influenced by a number of speaker and contextual factors, are always a subset of the whole, and are always changing.

I have also made a case for the need for generalizability of linguistic inferences for second language acquisition, language pedagogy, and teacher education purposes. In order for inferences to be useful for applied linguistics purposes, they have to address the "who," not only the "what" and they must be generalized at an optimal level depending on the purpose for the explanation. In addition, for pedagogical purposes, the level of generalizability should be negotiated between teacher and students so as to be neither too abstract nor too particularized for a given group of learners. Finally, I have pointed out that truly useful data would be those that have both psychological and linguistic validity for speakers and learners of a language.

References

Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language, 72*(1), 32–68.

Barlow, M. & Kemmer, S. (Eds.). (2000). *Usage based models of language*. Stanford, CA: CSLI Publications.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: CUP.

Bolinger, D. (1968). Entailment and the meaning of structures. *Glossa*, *2*(2), 119–127.

Breivik, L. (1981). On the interpretation of existential *there. Language, 57*(1), 1–25.

Celce-Murcia, M. (1980). Contextual analysis of English: Application to TESL. In D. Larsen-Freeman (Ed.), *Discourse analysis in second language research* (pp. 41–55). Rowley, MA: Newbury House.

Celce-Murcia, M. (1990). Data-based language analysis and TESL. In J. E. Alatis (Ed.), *Proceedings of Georgetown University roundtable on languages and linguistics 1990* (pp. 245–259). Washington, DC: Georgetown University Press.

Celce-Murcia, M. & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course* (2nd ed.). Boston: Heinle & Heinle.

Chomsky, N. (1995). *The minimalist program.* Cambridge, MA: The MIT Press.

Gass, S. & Mackey, A. (2000). *Stimulated recall methodology in second language research.* Mahwah, NJ: Lawrence Erlbaum.

Givón, T. (1980). The binding hierarchy and the typology of complements. *Studies in Language, 4*(3), 333–377.

Givón, T. (1999). Generativity and variation: The notion of "rule of grammar" revisited. In B. MacWhinney (Ed.), *The emergence of language* (pp. 81–114). Mahwah, NJ: Lawrence Erlbaum.

Hunston, S. & Francis, G. (2000). *Pattern grammar.* Amsterdam: John Benjamins.

Kennedy, G. (2003). Amplifier collocations in the British National Corpus: Implications for English language teaching. *TESOL Quarterly, 37*(3), 467–487.

Kim, K.-H. (1995). WH-clefts and left-dislocation in English conversation: Cases of topicalization. In P. Downing & M. Noonan (Eds.), *Word order in discourse* (pp. 245–296). Amsterdam: John Benjamins.

Knowles, P. (1979). Predicate markers: A new look at the English predicate system. *Cross Currents, VI*(2), 21–36.

Lakoff, G. (1987). *Women, fire, and dangerous things.* Chicago, IL: Chicago University Press.

Langacker, R. (1991). *Foundations of cognitive grammar.* Volume 2. Stanford, CA: Stanford University Press.

Larsen-Freeman, D. (1991). Teaching grammar. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (2nd ed., pp. 279–296). New York, NY: Newbury House/HarperCollins.

Larsen-Freeman, D. (1997). Chaos/complexity science and second language acquisition. *Applied Linguistics, 18*(2), 141–165.

Larsen-Freeman, D. (2000). Grammar: Rules and reasons working together. *ESL/EFL Magazine,* January/February, 10–12.

Larsen-Freeman, D. (2003). *Teaching language: From grammar to grammaring.* Boston, MA: Heinle & Heinle.

Larsen-Freeman, D. (2004). Enhancing contextual analysis through the use of linguistic corpora. In J. Frodesen & C. Holten (Eds.), *The power of context in language teaching and learning.* Boston, MA: Heinle & Heinle.

Larsen-Freeman, D., Kuehn, T., & Haccius, M. (2002). Helping students make appropriate English verb tense-aspect choices. *TESOL Journal, 11*(4), 3–9.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY: W. H. Freeman and Company.

Michigan Corpus of Spoken Academic English. English Language Institute. University of Michigan, Ann Arbor. Website: www.hti.umich.edu/m/micase

Newmeyer, F. (2003). Grammar is grammar and usage is usage. Presidential address delivered at the Annual Meeting of the Linguistic Society of America, Atlanta, January.

Rutherford, W. (1983). Language typology and language transfer. In S. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 358–370). Rowley, MA: Newbury House Publishers.

Sasaki, M. (1990). Topic prominence in Japanese EFL students' existential constructions. *Language Learning, 40*(3), 333–368.

Schachter, J. & Rutherford, W. (1979). Discourse function and language transfer. *Working Papers in Bilingualism, 19*, 1–12.

Whitehead, A. N. (1929). *The aims of education.* New York, NY: MacMillan.
Widdowson, H. (1990). *Aspects of language teaching.* Oxford: OUP.
Williams, H. (1996). An analysis of conjunctive adverbial expressions in English. PhD Dissertation. University of California, Los Angeles.
Yip, V. (1995). *Interlanguage and learnability: From Chinese to English.* Amsterdam: John Benjamins.

A conversation analytic perspective on the role of quantification and generalizability in second language acquisition

Numa Markee
University of Illinois at Urbana-Champaign

This chapter develops a methodological critique of quantitative, experimental approaches to input and interaction in mainstream, cognitive SLA from the qualitative perspective of ethnomethodological conversation-analysis-for-second-language-acquisition (CA-for-SLA). The chapter illustrates the substantive and methodological insights that may be gained from using a single, deviant case analysis approach to understand how language learning behavior is organized. This analysis also highlights issues such as the role of inferencing in the interpretation of data and problematizes the extent to which mainstream SLA studies are in a position to make valid generalizations about the function and organization of repair in language learning.

Introduction

I begin by reviewing the literature on the Interaction Hypothesis in second language acquisition (SLA) studies and then develop a methodological critique of experimental approaches to SLA on the basis of insights drawn from the field of conversation analysis (CA). The empirical analysis that follows of a learner's use of a rare, possibly unique, example of a particular type of Counter-Question in SL classroom talk highlights some of the substantive and methodological insights that may be gained from using a single case analysis approach to "conversation-analysis-for-second-language-acquisition" (CA-for-SLA) (Markee 2005). This analysis also highlights issues such as the role of inferencing in the interpretation of data and problematizes the extent to which SLA studies are in a position to make valid generalizations about the function and organization of repair in language learning. I conclude by outlining

what the prospects are for a grounded experimental approach to the study of repair in SLA.

The interactionist hypothesis: An overview

This chapter aims to offer a social constructivist (and hopefully constructive) critique of how we conceptualize, and do, SLA research on social interaction. Let me first acknowledge the depth, breadth, and dynamism of what – for want of a better description – I will call "mainstream" SLA studies during the past 25 years. For present purposes, the story begins with Hatch (1978), who argued that "one learns how to do conversation, one learns how to interact verbally and out of this interaction syntactic structures are developed" (p. 404). This remarkable statement of the Discourse Hypothesis (see also the work of Peck 1978, 1980; Sato 1986, 1988) continues to influence SLA studies to this day. More specifically, drawing on Hatch's ideas, Krashen (1980, 1981, 1982, 1985) developed the first theory of SLA (Monitor Theory), whose most important and enduring tenet was the Input Hypothesis. That is, Krashen suggested that learners learn SLs by being exposed to language that is slightly beyond their current level of competence – so called "*comprehensible input*" or "*i+1*."

In Krashen's work, i+1 is a static concept: it washes over learners, who pick up new language from contextual clues. However, in Long's work on the Interaction Hypothesis, comprehensible input is something that learners actively have to get for themselves (1980, 1981, 1983a, 1983b, 1985a). They do this by initiating a variety of conversational repairs with their native speaker (NS) or non-native speaker (NNS) interlocutors (Long 1983a; Long 1983b; Varonis & Gass 1985a, 1985b). Repair categories include comprehension checks, clarification requests, confirmation checks, and a variety of less commonly used categories, such as verifications of meaning, definition requests, and expressions of lexical uncertainty (Porter 1986). The function of repairs is to make initially incomprehensible talk progressively more understandable to learners as they attempt to negotiate meaning (Pica, Doughty, & Young 1986).

Repairs occur more frequently in NS-NNS talk than in native speaker-native speaker (NS-NS) interaction (Ellis 1985; Long 1980, 1981b, 1983b; Pica & Doughty 1985) and more frequently still in non-native speaker- non-native speaker (NNS-NNS) talk (Long & Porter 1985). Furthermore, two-way tasks promote more repairs than one-way tasks (Doughty & Pica 1986; Long 1980; Pica 1987). Convergent tasks also trigger more repairs than divergent tasks (Duff 1986). Moreover, unfamiliar tasks promote more repairs than familiar

tasks, and unfamiliar interlocutors repair their talk more often than familiar conversational partners do (Gass & Varonis 1984). In addition, more repairs occur in groups made up of people from different L1 backgrounds (Varonis & Gass 1985b). Mixed-gender groups engage in more repairs than same-gender groups (Gass & Varonis 1986; Pica, Holliday, Lewis, & Morgenthaler 1989. See also Long 1989, 1990; Long & Porter 1985; Pica 1992 for reviews). And more repairs occur in groups made up of mixed proficiency levels as opposed to groups in which learners are of the same proficiency (Yule & McDonald 1990). Note also that classroom-oriented research by Foster (1998) suggests that the occurrence of repairs may not be a function of task type so much as whether learners are put into pairs rather than small groups. According to this latter scenario, there is a great deal of variation at the individual level on whether repairs are initiated at all. When repair initiations do occur, they seem to occur more frequently in dyads than in small groups.

The opportunity to plan also seems to lead to greater negotiation (Crookes 1989; Foster & Skehan 1996; Mehnert 1998). Furthermore, task complexity, structure and processing load all have an impact on learners' performance (P. Robinson 1995; Skehan & Foster 1997). This body of research, combined with related work on language produced during small group work as opposed to lockstep work (Bygate 1988; Pica & Doughty 1985; Long, Adams, McLean, & Castaños 1977) has culminated in various empirically-based proposals for task-based language teaching (Long 1985b, 1989, 1991; Long & Crookes 1992, 1993; Nunan 1993; Pica, Kanagy, & Falodun 1993).

A key extension of the Input Hypothesis has been proposed by Swain (1985), who argues that learners also need to produce *comprehensible output* in order to move on from merely getting the semantic gist of what is being said to producing new language that is syntactically analyzed. Krashen's rebuttal of the output hypothesis notwithstanding (Krashen 1989), this position has received considerable empirical support in recent years (see Carroll & Swain 1993; Gass & Varonis 1994; Kowal & Swain 1994; Pica 1987, 1992; Pica, Holliday, Lewis, & Morgenthaler 1989; Shehadeh 1999; Swain & Lapkin 1995). Comprehensible output is currently thought to serve three main functions: (1) it promotes the "noticing" of new linguistic forms by learners; (2) it enables learners to test hypotheses about how the SL works; and (3) it also serves the metalinguistic function of allowing learners to control and internalize linguistic knowledge (Swain 1995).

In its latest version, the Interaction Hypothesis (IH) focuses on the role of attention, awareness, a focus on form, and the function of negative feedback in SLA:

> It is proposed that environmental contributions to acquisition are mediated by selective attention and the learner's developing L2 processing capacity, and that these resources are brought together most usefully, though not exclusively, during *negotiation for meaning*. Negative feedback obtained through negotiation work or elsewhere may be facilitative of L2 development, at least for vocabulary, morphology, and language-specific syntax, and essential for learning certain specifiable L1-L2 contrasts. (Long 1996: 414, emphasis in the original)

More specifically, contrary to the position adopted by Krashen (1985, 1989) and VanPatten (1988) that learning is a sub-conscious process, Kormos (2001), Schmidt (1990, 1993, 1994), and Schmidt and Frota (1986) suggest that adult SL learners must consciously notice or "apperceive" (Gass 1988, 1997) new language forms in the input in order for it to become available for learning. Support for this position is provided by theoretical and empirical studies on enhanced input (Doughty 1991; Sharwood-Smith 1991, 1993). This work suggests that subjects who focus on both form and meaning do better than learners who focus on isolated grammatical forms. Support for this position is also provided by classroom research (see, for example, Harley 1989; Lightbown & Spada 1990; Tomasello & Herron 1988; White, Spada Lightbown, & Ranta 1991) on the relative merits of what Long (1988, 1991) has called a focus on *form* (= learners paying attention to linguistic form in the process of engaging in meaning-oriented talk) rather than *forms* (= learners working on language as a decontextualized system; see also Doughty & Williams 1998; Muranoi 2000; Spada 1997; Williams 1999).

The role of *negative feedback* as a possible factor in second language (SL) learning has also become an important issue in SLA studies. Arguing from a Universal Grammar (UG) perspective, Gregg (1984, 1993, 1996) and White (1987, 1989, 1991) have cast doubt on the theoretical importance of i+1 in SLA. Indeed, White suggests that what learners need is not so much comprehensible input as *in*comprehensible input, and that positive evidence alone cannot serve as the sole means of destabilizing learner's interlanguage. Citing the example of different adverb placement rules in French and English, respectively, White (1991) argues that NSs of French learning English as a SL will not encounter any input in English that specifically prohibits verb-adverb-direct object strings (for example, "Je bois toujours du café" = literal translation: "I drink always coffee" = translation into standard English: "I always drink coffee"). SL learners thus need negative evidence to tell them that such a construction does not work in English (see also Birdsong 1989).

Research on the function and efficacy of negative feedback began with work on the effects of error correction on learner output (Brock, Crookes, Day,

& Long 1986; Chaudron 1977, 1987, 1988; Crookes & Rulon 1988; Lightbown & Spada 1990; Salica 1981; Spada & Lightbown 1993; White 1991; White, Spada, Lightbown, & Ranta 1991). Later work has tended to focus on the relative effectiveness of different types of implicit and explicit negative feedback, particularly recasts and reformulations of the input (Doughty & Varela 1998; Gass 1997; Long, Inagaki, & Ortega 1998; Lyster 1998; Lyster & Ranta 1997; Mackey & Philp 1998; Nicholas, Lightbown, & Spada 2001; Oliver 1995, 1998, 2000). While the jury is still out on the precise role played by negative feedback in SLA, it is likely that such feedback facilitates SL leaning, and may also be necessary for learning some L2 structures (Long 1996).

A critique of mainstream work on SLA

As I have already noted, the body of research spawned by the IH is impressive. And yet . . . there are various epistemological, methodological and substantive issues that remain unanswered. For present purposes, I couch my discussion of the issues in terms of Schegloff's (1993) critique of attempts to quantify social interaction.

Despite the noteworthy contributions of Hawkins (1985) and Swain and her associates to the formulation of the IH, qualitative research is dramatically under-represented in SLA studies. The scarcity of CA studies is especially noticeable here, because the concept of repair is borrowed from CA (see Jefferson 1987; Schegloff 1979, 1991, 1992, 1997a, 1997b, 2000; Schegloff, Jefferson, & Sacks 1977). However, this situation is now beginning to change. Two early studies by Gaskill (1980) and Schwartz (1980) on SL repair have been followed up in the last few years by a spate of CA-for-SLA work (see Firth & Wagner 1997; Kasper 2002; Kasper & Ross 2001; Markee 1994, 1995, 2000, 2003, 2004a, 2004b, 2005a, 2005b; Mori 2002; Seedhouse 1997, 1999; Wagner 1996; Willey 2001). Other writers have also used CA techniques – for example, van Lier (1988), or, more recently, Lazaraton (2003, 2004), who labels her work "microanalysis" – but do not claim to be doing CA *per se*. CA techniques have also been used by researchers who frame their work in terms of sociocultural theory (Ohta 2001a, 2001b), systemic grammar (Young & Nguyen 2002), or, potentially, variationist approaches to SLA (Tarone & Liu 1995). Finally, the work of Koshik (2002a, 2002b, 2003); Lerner (1995), McHoul (1978, 1990) and Olsher (2001) on the structure of classroom discourse is also relevant to CA-for-SLA research.

These developments are an encouraging step in the right direction. But this body of research is still dwarfed, both in size and influence, by a longer, better established and, above all, a predominantly *experimental*, tradition in mainstream SLA. Now, CA-for-SLA does not have an inalienable right to have a seat at the SLA table: it has to demonstrate that its insights are relevant and useful to SLA studies. Some of the pertinent issues have already been vigorously debated (see the exchanges between Firth & Wagner 1997 on the one hand, and Gass 1998, 2001; Kasper 1997 and Long 1997, 1998 on the other). Here, I review the issues in terms of the *numerator*, *denominator*, *significance,* and *domain* problems identified by Schegloff (1993), which all affect the possibility of meaningful quantification of social interaction in SLA studies.

The domain problem

It is a fundamental tenet of the IH that "free conversation is notoriously poor as a context for driving interlanguage development . . . in contrast, tasks that orient participants to shared goals and involve them in some work or activity produce more negotiation work' (Long 1996: 448). Let me now restate this claim in CA-for-SLA terms to highlight certain problems: (1) Ordinary conversation and institutional talk (specifically, classroom talk) are observably different speech exchange systems; (2) classroom talk provides better structural opportunities for SLA to occur than ordinary conversation does because the repair practices that participants orient to as they do classroom talk provide qualitatively better opportunities for learners to notice new forms and to negotiate meaning; (3) classroom talk provides a greater number of opportunities than ordinary conversation does for learners to engage in such noticing and negotiation. I fully agree with the first of these propositions and have called for just such a research agenda (Markee 2000). CA originally focused on the study of ordinary conversation, which may be glossed as "casual, social talk that routinely occurs between friends and acquaintances, either face-to-face or on the telephone" (Markee 2000: 24). Over time, it has also come to encompass the study of institutional talk such as news, medical, courtroom and classroom talk (see, for example, Boden & Zimmerman 1991; Button 1991; Clayman & Heritage 2002; Drew & Heritage 1992; Heath 1989; Heritage & Roth 1995; McHoul 1978, 1990; J. D. Robinson 1998; Stivers 2001). CA clearly possesses both the methodological tools and the expertise that are necessary to explicate how speech exchange systems differ one from another.

I am also fascinated by Propositions 2 and 3. However, mainstream SLA and CA-for-SLA *both* have to confront seven unresolved problems if the im-

plications of these three propositions are to be sustained: (1) Despite the vast amount of experimental research that has been done on the role and function of negotiation in mainstream SLA, few benchmark qualitative studies exist that systematically compare and contrast the sequential, turn-taking and repair practices of (SL) ordinary conversation with those of (SL) classroom talk (Liddicoate 1997; however, see also Kasper 2002). (2) This lack of benchmarks is highly problematic for mainstream SLA, because Proposition 1 is the theoretical foundation on which Propositions 2 and 3 rest. (3) At the moment, mainstream SLA research does not adequately distinguish between ordinary conversation and institutional talk, nor among different institutional varieties of talk (see, for example, Varonis & Gass 1985b). (4) Mainstream SLA studies seems curiously uninterested in documenting and analyzing the observable learning consequences of specific conversational acts by learners. Thus, while the theoretical construct of i+1 may eventually yield interesting insights about SLA as an abstract, aggregated, group phenomenon, we are still rarely offered analyses that show how individual learners actually do comprehensible input in real time, and how such input leads first to understanding and then to learning that is instantiated as comprehensible output, if only in the short term (for an example of such research, Markee 1994, 2000). (5) Until we have a better qualitative understanding of how these speech exchange systems (i.e., *domains*) are organized, continuing attempts to quantify talk-in-interaction and to generalize from these data will inevitably be premature. To be valid and reliable, experimental work must be properly grounded in prior analyses that explicate *how that particular piece of talk* was produced *at that particular moment* in *that particular speech event* to achieve *that particular action.* (6) Whatever results of the IH are eventually sustained, current claims concerning the role of repair in language learning posited are likely only generalizable to the domain of *instructed* SLA, not SLA *as a whole*. Finally, note that the idea encapsulated in Proposition 1 goes far beyond issues of quantification. It also invokes one of the most vexing controversies in SLA studies today.

I am referring, of course, to the question of how psycholinguistic questions of language learning intersect with sociolinguistic aspects of language use. As Long (1997, 1998) correctly notes in his responses to Firth and Wagner (1997), social cognitive researchers working within the framework of the IH have conducted a great deal of research on the role of conversational repair in facilitating negotiation. However, in the face of radical, sociolinguistically-inspired critiques of their work by Firth and Wagner (1997) and others, there seems to be a tendency among some prominent social cognitivists to backpedal

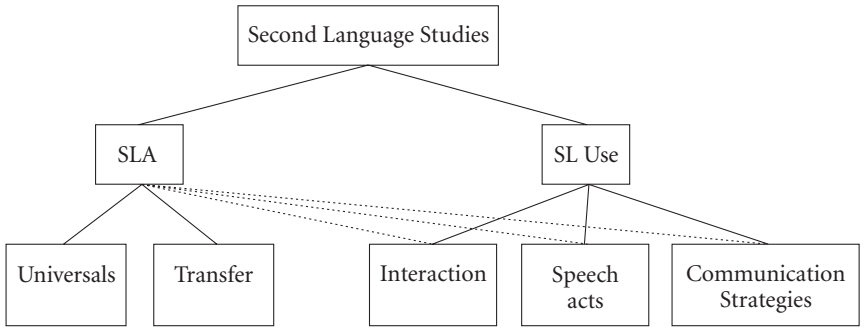

Figure 1. A characterization of research in "SLA" (Gass 1998: 88)

from the logical implications of their earlier work and to retreat into a world of psycholinguistic isolationism. For example, Gass asserts that:

> . . . it is true that in order to examine [changes in linguistic knowledge], one must consider language use in context. But in some sense this is trivial; the emphasis in input and interaction studies is on the *language* used and not on the act of communication. (Gass 1998: 84; emphasis in the original)

This position maintains an untenable distinction between psycholinguistic and sociolinguistic views of language. Of course, Gass is sensitive to such a criticism and softens her position by saying that "views of language that consider language as a social phenomenon and views of language that consider language to reside in the individual do not necessarily have to be incompatible" (Gass 1998: 88). Unfortunately, this turns out to be little more than a *pro forma* disclaimer, because Gass specifies the relationship between studies in SLA and SL use as shown in Figure 1.

Gass finds some of Firth and Wagner's arguments regarding the scope of SLA puzzling, but her attempts to compartmentalize issues of language use and language acquisition illustrated in Figure 1 are equally odd. An empirically grounded understanding of how learners' interlanguage knowledge (as this is reflected *in* and *through* their talk-in-interaction) progresses from A to B, and what "events promote or hinder such progress" (Kasper 1997: 310) cannot be dismissed as a "trivial" issue. It is a crucial foundation for the IH. If this means that advocates of the IH have to accept that language acquisition and use are indivisible components of the SLA enterprise, then this is not to be seen as a threat to the disciplinary integrity of SLA studies. It is a consequence of the IH's own theoretical interests in social interaction as a resource for SLA. I return to these issues in the empirical and concluding sections of this chapter.

The significance problem

A key assumption of experimental research is that the nature or strength of relationships or claimed differences between treatment effects, etc. must be *significant*, in the technical, statistical sense of this word. That is, in order to develop a viable mathematical model of hypothesized relationships between independent and dependent variables, we must be able to find enough *aggregated instances* of a given behavior to be confident that a finding is robust. However as Schegloff (1993) points out, this etic, or researcher's, perspective on knowledge construction is not the only way of conceptualizing the *relevance* of analytical findings. As Schegloff (1993: 101) memorably declares, "*one* is also a number" (emphasis in the original). What he means by this is that, from an emic, or participant's, perspective:

> The best evidence that some practice of talk-in-interaction does, or *can* do, some claimed action, for example, is that some recipient on some occasion shows himself or herself to have understood it, most commonly by so treating it in the ensuing moments of the interaction, and most commonly of all, next. Even if no quantitative evidence can be mustered for a linkage between that practice and that resultant "effect," the treatment of the linkage as relevant – by the *parties* on *that* occasion, on which it *was* manifested – remains.
>
> (Schegloff 1993: 101, emphasis in the original)

This position has at least two profound implications for the way in which CA research is carried out. First, the *warrant* for any analytic claims that are made about how ordinary conversation and institutional talk are organized must be located in the local context of participants' talk, and explicated in terms of what members understand each other to be doing at the time that they are doing it. Thus, CA work is always based in the first instance on the exhaustive, micro-analysis of *single cases* of interaction. Second, analyses of *collections* (= "aggregates of single instances" of, say, second position repairs; see Schegloff 1993: 102) may also be carried out. However, since the goal of analysts is to develop a member's understanding of participants' behaviors, there can be no such thing as an "outlier," which may be discarded as unrepresentative of group norms. In order to account for apparently deviant behaviors, *deviant case analysis* is used to provide a complete account of a phenomenon. The most well known example of this technique is Schegloff's (1968) analysis of sequencing in conversational openings on the telephone. His first analysis accounted for 499 out of 500 cases in the collection. However, Schegloff reanalyzed the entire corpus to yield all 500 cases of this phenomenon. I return to these issues in the empirical section of this chapter.

These issues are generally not well understood in mainstream SLA. For example, the fact that most CA research does not attempt to present data in a quantified form is seen as an annoying, almost irrational quirk by some writers (see Long 1997), rather than as the product of a principled epistemological stance. Furthermore, in the rush to quantify data that has characterized mainstream SLA studies from its inception, the fact that experimental research in SLA is not without its own problems has largely been ignored (see Markee 2000: Chapter 2).

First, the functional categories of repair used by mainstream SLA research (comprehension checks, clarification requests, confirmation checks, recasts, etc.) are frequently so ambiguous or decontextualized that it is often not clear whether a particular fragment of talk actually constitutes, say, a comprehension check, or a completely different category of repair (Ohta 2001b). The ambiguous nature of these categories is problematic from an emic perspective because it is not clear that members orient to these categories as distinct constructs. And from an etic perspective, the decontextualization of these categories is even more problematic, since experimentation requires that categories used for coding be discrete in order for the subsequent analysis to be meaningful. Varonis and Gass (1985b) have acknowledged this problem (see also how Oliver 1998, 2000 deals with double-coding issues) but this admission does not seem to have dampened the enthusiasm with which these categories are still employed in experimental studies.[1]

Second, mainstream SLA's preoccupation with quantifying SL data as the default mode of analysis has prompted Aston (1986) to point out that a quantitative, "more the merrier" approach to investigating SL repair fails to acknowledge that there may be considerable negative social consequences for members who engage in *excessive* repair of their interlocutors' talk.

The denominator problem

Schegloff (1993) explains what the denominator problem is by critiquing a quasi-experimental study that sought to quantify the notion of sociability in terms of how many times *per minute* subjects laughed during the experiment. After noting that laughter in naturally occurring talk-in-interaction is always a *responsive* phenomenon, whose quality and placement in the ongoing talk therefore matter in terms of members' assessments of whether such laughter is affectively appropriate or not, he points out:

> If one wants to assess how much someone laughed, to compare it with other laughter by that person or by others, then a denominator will be needed that is *analytically relevant to what is to be counted* because it is *organizationally related to it in the conduct of interaction*. And minutes are not.
>
> (Schegloff 1993: 104, emphasis in the original)

This suggests that work on repair in SLA that quantifies how often repairs occur without specifying *a relevant denominator* (for example, the number of repairs that occur *per task type*) are likely to be premature. This is because the specification of task as a denominator requires prior grounded research to establish whether, for example, one-way and two-way tasks constitute distinct domains. To date, this research has not been carried out in mainstream SLA studies. I return to this issue in the conclusion to this chapter.

The numerator problem

The numerator problem has to do with the raw frequency of a particular behavior in talk. So, to continue our repair example, quantifying the number of repairs that occur in a speech event or aggregation of speech events to find out whether this number reaches statistical significance does not tell us anything about *how* repair is achieved by participants *in* and *through* talk. If quantified data are to be meaningful, we need to understand that the observable *absence* or *rarity* of specific types of repair is as pertinent to a comprehensive analysis of repair as their *expected presence* or *frequency* in talk. We are in fact *required* to develop the notion of an "*environment of relevant possible occurrence*" (Schegloff 1993: 106). So, we must have a detailed sequential understanding of *where and when an analytic warrant exists* for saying that a repair is present, absent, frequent, or rare in a given piece of talk. Again, we can only do this by grounding our analyses of an individual case in the practices that distinguish one speech exchange system from another. I illustrate how to do this in the following section.

Conversation analysis: An emprical example

In the empirical analysis that follows, as required by Kasper (1997), I propose to show that CA can explicate how certain events in classroom talk can hinder or at least delay acquisitionally relevant talk. Furthermore, to connect again with the theme of CA's stance on quantification and generalizability examined

earlier in this paper, I make this argument by using deviant case analysis to interpret one student's use of a rare, possibly unique, example of a Counter-Question – that is, a type of question that is inserted, normally by teachers but in this case by the student – between the first and second pair parts of a Question-Answer adjacency pair. Of course, from an experimental perspective, such data would never be used to make generalizations about SLA. But from a CA point of view, it is the very *rarity* of the example that allows us to develop a deeper *strategic* understanding of the different rights and responsibilities of members in teacher-student speech exchange systems. It is these deeper strategic insights about the structural organization of talk that are potentially generalizable, not the specific *tactics* of individual participants in a particular conversation.

Fragment 1 below (see Markee 1995 for a full analysis) comes from a 50 minute undergraduate university ESL class that was video- and audio-taped in 1990. The class, which was discussing the issue of German reunification, was taught by an experienced NS English teacher. The methodology used by the teacher involved task-based interactions mediated through small group work, which provided learners with opportunities to focus on form on an as-needed basis.

During dyadic talk between L9 and L11 that occurred before the interaction reproduced below, L11 had had trouble on two separate occasions trying to work out what the phrase "you pretend to pay us and we pretend to work" means. Note that this prior talk was constructed as a speech exchange system that approximated in many ways the locally organized turn-taking and repair practices of ordinary conversation (Sacks, Schegloff, & Jefferson 1974). That is, although the topics of L9's and L11's talk were preset by comprehension questions in the materials, neither participant had a pre-allocated right to take or assign turns or to initiate repairs. However, when we join the interaction at line 09 of Fragment 1, where L9 seeks the teacher's (= Jane's) help, we are witnessing the beginning of L9 and L11's third attempt to resolve this problem (see Appendix 1 for transcription conventions).

Fragment 1

((*L9 and L11 are looking down at their reading materials. L9 is holding the pages of his materials in his right hand, and L11 is leaning his head on his left hand.*))

01	L9:	[((*L9 leans toward L11*))]
02	L9:	can [we call jane maybe, ((*unintelligible*)).
03	L9:	[X_____.....X____
04		(0.3)
05	L11:	myeah.

[((*L9 and L11 both look up toward the front of the class. L9 holds his chin in his right hand in a thinking posture, and L11 rests his left hand on his left thigh*))]
L9: Q [nt jane?]
T: uh huh?
(-------[---) ((*T's shadow on the floor indicates she is beginning to move toward L9 and L11*))
L11: [X_____
L9: Q [your input plea[h huh [huh] huh] ((*L9 smiles as he begins to laugh and looks down during the last two laughter tokens*))
T: [huh]
L11: [h huh [huh]huh] [huh. huh
L11: _____
((*T appears in the camera shot. She is approaching L9 and L11*))
L11: _____
L9: Q ·hhhhh there is this e:::h ((*diminishing volume through "e::::h"*))
((*L9 looks down at his reading materials*))
(0.6) some sort of an idiom
((*T arrives next to L9 and L11*))
Q you pretend to pay us [and we pretend to work
[((*T leans down and reaches with her right hand to identify where the problem phrase is in L9's reading text*))
[((*L9 and L11 both look up at T*))
T: CQ(D) [ok. [what do you think that could be:]
[((*T stands up straight as she speaks.*))
CQ(D) (0.3) do you have any idea? ((*T is looking at L11. She holds the open palm of her right hand up and makes a slight beating gesture as she says the word "you".*))
L9: [X_____((*L9 looks at L11*))_____
L11: CQ → [do you- do you know what the word pretend, (.) means,]
[((*L11 raises his left hand to touch his his nose and beats out the stress on the words "know" and "pretend". His hand then comes to a resting, thinking posture position, supporting his chin. As she listens to L11, T holds her chin in the fingers of her right hand in a thinking posture.*))
(1.0)
T: CQ do [I know what the word pretend means]
[((*T leans forward slightly toward L11 and touches her chest with her right hand as she says the word "I". In addition, her intonation steadily rises through the turn, ending on a high note. Meanwhile,*

48 *L9 briefly looks at T as she says the word "I" and then moves his*
49 *gaze back onto L11.*]))
50 L11: A [yeah- I- I /dawt/-]
51 [((*L11 does a circling gesture in front of him with his left hand, which*
52 *ends with his open left hand resting on his chest as he says the second*
53 *"I"*))]
54 L11: [I don't know that see]
55 [((*L11 shakes his head slightly four times*))]
56 T: CQ(D) oh ok °who-° °do-° ((*T looks quickly to her right and then her left at*
57 *the rest of the class.*)) does anybody know what the word pretend
58 means. ((*T is speaking to the whole class.*))

At lines 09, 11, and 14–26, L9 calls the teacher over and identifies the phrase "you pretend to pay us and we pretend to work" as an idiom that he and L11 do not understand. At lines 31 and 33, instead of answering this question directly, T does a Counter-Question (CQ) turn that is formulated as a display (D) question. This verbal behavior (along with the accompanying gestural behaviors at lines 33–35 that clearly nominate L11 as next speaker) may be analyzed as a move that puts T back in sequential position not only to *ask* questions to which there is a known answer but also to *comment* on students' answers in the commenting C slot of Question-Answer-Comment (QAC) sequences (see McHoul 1978, and Markee 1995, on QAC sequences and also the earlier work of Mehan 1979, and Sinclair & Coulthard 1975, on Initiation-Response-Feedback sequences in classroom talk). Thus, by initiating this CQ(D) turn at lines 31 and 33, T asserts her instructor's right to control the pedagogical agenda by orienting to the practices of classroom talk, not ordinary conversation. The prototypical trajectory of CQ(D) sequences may therefore be summarized as follows:

Owner of the turn:	L →	T →	L	T
Turn type	Q →	CQ(D) →	A	C

Figure 2. Trajectory of CQ(D) sequences (Markee 1995: 75)

But, at line 37, L11 asks whether T knows what the word "pretend" means. Now, L11 does not stress the word "you" here. He may be trying to tell T that it is *he*, not *L9*, who does not understand the meaning of this word, a piece of information which T does not know, since she has not participated in L9 and L11's prior dyadic talk. Whatever L11 may have been trying to accomplish at line 37, T reacts at line 44 by treating L11's turn as a CQ turn that challenges her status as a NS English teacher. She accomplishes this by indicating that there

is potentially some trouble at line 43, where she pauses for a full second. She then does a CQ turn of her own at line 44 that rhetorically demands recognition of her status as a NS teacher as the preferred response (note T's use of the heavy contrastive stress on the word "I", the rising intonation through the turn at line 44, and the accompanying, highly emphatic, visual deixis of the hand gesture at line 45). At line 50, L11 begins by reiterating his question by saying "yeah" (a dispreferred, or marked, response). However, after some initial perturbations (note the cut-offs in the first part of this turn), he displays a new understanding of what the teacher is doing in her previous turn and begins to repair his social relationship with T. His first attempt still comes out rather garbled: "I- I [dawt]-" is not only marked by hesitations and cut-offs, but the word phonetically transcribed as [dawt] may be a first attempt to say "I don't know that", which seems to "come out wrong" under the communicative stress of the moment. In the completion of this turn at line 54, L11 repairs this first attempt by saying "I don't know that see" and achieves the preferred response of acknowledging that he does not know what the word "pretend" means and that he therefore needs T's expert help.

At line 56, T first acknowledges her new understanding and acceptance of L11's clarification by using the change of state token "oh ok" (Heritage 1984) and then, as in lines 31 and 33, moves on in the second part of her turn to redirect the question to other interlocutors, in this case, the rest of the class. This Q turn is again done as a D question, and the rest of the of the talk (not reproduced here) runs off in the canonical QAC order. The actual trajectory of Fragment 1 can thus be diagrammed as follows:

Owner of the turn:	L	T	L	T	L	T
Turn type	Q	CQ(D)	CQ	CQ	A	Q(D) …

Figure 3. Trajectory of Excerpt 1

From a CA-for-SLA perspective, this excerpt is of considerable analytical interest. Although L11's behavior is unique in the eight classes and approximately 26 hours of classroom recordings that form my database, the analysis summarized in Figure 3 is not a counter-example to the analysis of the underlying practices that govern the speech exchange system extrapolated from the prototypical sequential trajectory shown in Figure 2. In fact, this deviant case analysis demonstrates that teachers have pre-allocated rights to doing specific turns (Q, CQ(D) and A) in QAC sequences and that learners do *not* have the right to do turns that may be interpreted as CQ(D) talk. Furthermore, if

a student unexpectedly fails to orient to the sequentially located turn-taking conventions and repair practices of classroom talk, this transgression is an act which the teacher may censure through the use of a rhetorical CQ(D) turn that other-initiates repair.

Note also that *how* participants resolve the ambiguity of *which* speech exchange system they are orienting to at *that* particular moment in time as they negotiate *that* particular repair is a complex issue. For example, the use of the change of state token "oh ok" by T at line 56 is notable because we know that this conversational object is characteristic of ordinary conversation, not institutional talk (Heritage 1984). This is because teachers rarely ask learners true information questions (Long & Sato 1983).[2] Furthermore, learners tend to ask teachers very few questions compared to the number of questions that instructors ask students (Dillon 1981, 1988; White & Lightbown 1984).

Here, T asks L11 at line 44 a question that is fishing for the preferred answer that T *does* know what the word "pretend" means. But the way L11 actually does this at lines 50 and 54 is by formulating his answer in terms of *his* ignorance concerning what "pretend" means, not the *teacher's*. T then responds at line 56 by saying "oh ok", thus exploiting a practice of ordinary conversation that hearably treats L11's answer as new information that she had not understood before. This verbal tactic enables T to end the confrontation and to align with L11's attempt to repair his social relationship with her.

Conclusion

I conclude this chapter with the following six observations. First, we should note that the disciplinary preference for explanation over interpretation and inference that has characterized applied linguistic and SLA research over the last 30 years is in the process of changing. For example, in their influential call for a re-opening of the research agenda on SL motivation, Crookes and Schmidt (1991) issued a call for more qualitative as well as quantitative research on the issues. Dörnyei (2000) and Dörnyei and Csizér (in press), whose own work is heavily experimental, have issued similar calls for more situated, process-oriented accounts of motivation. And McGroarty (1998) has gone even further, claiming that social constructivist approaches to theory building about language learning and use may provide the most interesting sources of insight for future applied linguistic research. Thus, the issues discussed here are embedded in a larger discussion of the potential value of quantification in such research.

Second, the empirical analysis of Fragment 1 responds to Kasper's (1997) call for CA-for-SLA to show how social events promote or, as in this case, *hinder* the possibility of interlanguage development from occurring. As we have seen, the social relationship between T and L11 had to be repaired before further language learning-oriented talk could continue. The *tactical* face-saving work done in Fragment 1 illustrates the larger *strategic* fact that classrooms are *social* environments as much as they are *learning* places. This conclusion also demonstrates how CA-for-SLA can *empirically* confront Gass' (1998) *theoretical* claim that language acquisition and language use are distinct aspects of second language studies. The clear implication of this analysis is that they are so closely intertwined as to be theoretically inseparable.

Third, despite the great importance that CA attaches to single case analysis, it is important to understand that CA does not *a priori* deny the value of a quantitative approach to social interaction. What I have argued in this chapter is that we must develop a rigorous, qualitative understanding of *how* SL learning and use are done by participants in order to motivate grounded quantitative research (for recent examples of such grounded experimental work in CA, see Heritage & Stivers 1999; Stivers 2001).

Fourth, in an ideal world, this grounding work would have been done *prior* to embarking on a significant program of experimental research. From a purely pragmatic point of view, we are clearly well beyond the point of being able to observe any so-called canonical order of doing qualitative research first and then following up with quantitative research (see also Crookes & Schmidt 1991 on this point). Thus, under present circumstances, CA-for-SLA research has to take on the unusual epistemological function of *confirming*, not just *generating* hypotheses about SL learning and use (Markee 2000).

Fifth, we must ultimately develop both interpretive and predictive explanations of SLA processes that are coherent in their own terms and that are also properly informed by each other. Indeed, as Schegloff (1993) suggests, the social construction of repair is a type of behavior that could in the long term potentially benefit from follow up experimental studies that have been properly grounded in qualitative CA work. But we are not there yet and, in terms of what the future holds for SLA studies, mainstream SLA researchers who use quantified data can expect to be asked with increasing insistence – as should qualitative researchers also: see Edge and Richards (1998) – to "show their warrant" for making claims X, Y or Z.

Finally, I accept that the development of a CA-for SLA agenda is controversial, in that it broadens the scope of mainstream SLA research, challenges conventional notions of the "proper" relationships between qualitative and

quantitative approaches to scholarship, and also forces us to rethink *what* and *how* we generalize from data. At the same time, I wish to argue that the results of such a respecification of our field will ultimately strengthen SLA studies, not weaken their fundamental disciplinary integrity. Proponents of different versions of SLA should certainly continue to engage in vigorous debate about the strengths and limitations of the research traditions within which we work. Ultimately, however, we are all concerned with explicating the same complex phenomenon of how and why SLs are learned, and it is in this cooperative spirit that the arguments developed in this paper have been offered.

Notes

1. In other words, this is a problem that cannot be solved by better inter-rater/coder reliability procedures.

2. Duff (personal communication, January 31, 2005), suggests that this is perhaps too strong a statement. While the use of true information questions was certainly rare in the early 1980s, there is increasing evidence that language teachers now do ask many more such questions today than they did 25 years ago.

Appendix 1: Transcription conventions

CA transcription conventions (based on Atkinson & Heritage 1984b).

Identity of speakers

T:	teacher
L1:	identified learner (Learner 1)
L:	unidentified learner
L3?:	probably Learner 3
LL:	several or all learners simultaneously

Simultaneous utterances

L1: [yes L2: [yeh	simultaneous, overlapping talk by two speakers
L1: [huh? [oh] I see] L2: [what] L3: [I dont get it]	simultaneous, overlapping talk by three (or more) speakers

Contiguous utterances

= (a) turn continues at the next identical symbol on the next line
(b) if inserted at the end of one speaker's turn and the beginning of the next speaker's adjacent turn, it indicates that there is no gap at all between the two turns

Intervals within and between utterances

(0.3) (1)	(0.3) = a pause of 0.3 second; (1.0) = a pause of one second.

Characteristics of speech delivery

?	rising intonation, not necessarily a question
!	strong emphasis, with falling intonation
yes.	a period indicates falling (final) intonation
so,	a comma indicates low-rising intonation suggesting continuation
go:::d	one or more colons indicate lengthening of the preceding sound; each additional colon represents a lengthening of one beat
no-	a hyphen indicates an abrupt cut-off, with level pitch
because	underlined type indicates marked stress
SYLVIA	capitals indicate increased volume
°the next thing°	degree sign indicates decreased volume
·hhh	in-drawn breath
hhh	laughter tokens

Commentary in the transcript

((coughs))	verbal description of actions noted in the transcript, including non-verbal actions
((unintelligible))	indicates a stretch of talk that is unintelligible to the analyst
. . . . (radio)	single parentheses indicate unclear or probable item

Eye gaze phenomena

The moment at which eye gaze is coordinated with speech is marked by an X and the duration of the eye gaze is indicated by a continuous line. Thus in the example below, the moment at which L11's eye gaze falls on L9 in line 412 coincides with the beginning of his turn at line 413

[X___((*L11's eye gaze is now directed at L9*))_____

412 L11: → ≈ [°**you can call me and then [I can say you] the the address**

413 L9: → [((*L9 nods 4 times*]

Eye gaze transition is shown by commas

[X_____, ,

The moment at which there ceases to be eye contact (as when a participant looks down or away from his/her interlocutor) is shown by periods

[X_____. . .

Other transcription symbols

co[l]al	brackets indicate phonetic transcription
→	an arrow in the margin of a transcript draws attention to a particular phenomenon the analyst wishes to discuss.

References

Aston, G. (1986). Trouble-shooting in interaction with learners: The more the merrier? *Applied Linguistics, 7,* 128–143.

Atkinson, J. M. & Heritage, J. (Eds.). (1984). *Structures of social action.* Cambridge: CUP.

Boden, D. & Zimmerman, D. H. (Eds.). (1991). *Talk and social structure.* Cambridge: Polity Press.

Brock, C., Crookes, G., Day, R., & Long, M. H. (1986). The differential effects of corrective feedback in native speaker/non-native speaker conversation. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 229–236). Cambridge, MA: Newbury House.

Button, G. (Ed.). (1991). *Ethnomethodology and the human sciences.* Cambridge: CUP.

Bygate, M. (1988). Units of oral expression and language learning in small group interaction. *Applied Linguistics, 9*, 59–82.

Caroll, S. & Swain, M. (1993). Explicit and implicit feedback. An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition, 15,* 357–386.

Chaudron, C. (1977). A descriptive model of discourse in the corrective treatment of learners' errors. *Language Learning, 27,* 29–46.

Chaudron, C. (1987). The role of error correction in second language teaching. In B. K. Das (Ed.), *Patterns of classroom interaction in South East Asia* (pp. 17–50). Singapore: SEAMEO-RELC Regional Language Centre.

Chaudron, C. (1988). *Second language classrooms. Research on teaching and learning.* Cambridge: CUP.

Clayman, S. & Heritage, J. (2002). Questioning Presidents: Journalistic deference and adversarialness in the press conferences of Eisenhower and Reagan. *Journal of Communication, 52*, 749–777.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition, 11*, 367–383.

Crookes, G. & Rulon, K. A. (1988). Topic and feedback in native speaker/non-native speaker conversation. *TESOL Quarterly, 22,* 675–681.

Crookes, G. & Schmidt, R. W. (1991). Motivation: Reopening the research agenda. *Language Learning, 41,* 469–512.

Dillon, J. T. (1981). Duration of response to teacher questions and statements. *Contemporary Educational Psychology, 6,* 1–11.

Dillon, J. T. (1988). The remedial status of student questioning. *Journal of Curriculum Studies, 20,* 197–210.

Dörnyei, Z. (2000). *Motivation*. London: Longman.

Dörnyei, Z. & Csizér, K. (2002). Some dynamics of language attitudes and motivation: Results of a longitudinal nationwide survey. *Applied Linguistics, 23,* 421–462.

Doughty, C. (1991). Second language instruction does make a difference. Evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition, 13,* 431–469

Doughty, C. & Pica, T. (1986). "Information gap" tasks: An aid to second language acquisition? *TESOL Quarterly, 20*, 305–325.

Doughty, C. & Varela, C. (1998). Communicative focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom language acquisition.* Cambridge: CUP.
Doughty, C. & Williams, J. (Eds.). (1998). *Focus on form in classroom language acquisition.* Cambridge: CUP.
Drew. P. & Heritage, J. (Eds.). (1992a). *Talk at work: Interaction in institutional settings.* Cambridge: CUP.
Duff, P. A. (1986). Another look at interlanguage talk: Taking task to task. In R. Day (Ed.), *Talking to learn* (pp. 147–181). Rowley, MA: Newbury House.
Edge, J. & Richards, K. (1998). May I see your warrant please? Justifying outcomes in qualitative research. *Applied Linguistics, 19*, 334–356.
Ellis, R. (1985). Teacher-pupil interaction in second language development. In S. M. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 69–85). Rowley, MA: Newbury House.
Firth, A. & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *The Modern Language Journal, 81,* 285–300.
Foster, P. (1998). A classroom perspective on the negotiation of meaning. *Applied Linguistics, 19,* 1–23.
Foster, P. & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition, 18*, 299–324.
Gaskill, W. (1980). Correction in native speaker – non-native speaker conversation. In D. Larsen-Freeman (Ed.), *Discourse analysis in second language research* (pp. 125–137). Rowley, MA: Newbury House.
Gass, S. M. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics, 9*, 198–217.
Gass, S. M. (1997). *Input, interaction and the second language learner.* Mahwah, NJ: Lawrence Erlbaum.
Gass, S. M. (1998). Apples and oranges: Or, why apples are not oranges and don't need to be. A response to Firth and Wagner. *The Modern Language Journal, 82,* 83–90.
Gass, S. M. (2001). SLA: A 2001 space odyssey. Plenary address to the Annual AAAL Conference, St. Louis, MI, March 2001.
Gass S. M. & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of non-native speech. *Language Learning, 34*, 65–89.
Gass, S. M. & Varonis, E. M. (1986). Sex differences and in non-native speaker/native speaker interaction. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 327–351). Cambridge, MA: Newbury House.
Gass, S. M. & Varonis, E. M. (1994). Input, interaction and second language production. *Studies in Second Language Acquisition, 16,* 283–302.
Gregg, K. R. (1984). Krashen's monitor and Occam's razor. *Applied Linguistics, 5,* 79–100.
Gregg, K. R. (1993). Taking explanation seriously; or, let a couple of flowers bloom. *Applied Linguistics, 14,* 276–294.
Gregg, K. R. (1996). The logical and developmental problems of second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 49–81). New York: Academic Press.
Harley, B. (1989). Functional grammar in French immersion: A classroom experiment. *Applied Linguistics, 10*, 331–359.

Hatch, E. M. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 402–435). Rowley, MA: Newbury House.

Hawkins, B. (1985). Is an 'appropriate' response always so appropriate? In S. M. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp.162–178). Rowley, MA: Newbury House.

Heath, C. (1989). Pain talk: The expression of suffering in the medical consultation. *Social Psychology Quarterly, 52,* 113–125.

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In J. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 299–345). Cambridge: CUP & Paris: Editions de la Maison des Sciences de l'Homme.

Heritage, J. & Roth, A. L. (1995). Grammar and institution: Questions and questioning in the broadcast news interview. *Research on Language and Social Interaction, 28,* 1–60.

Jefferson, G. (1987). On exposed and embedded correction in conversation. In G. Button & R. E. Lee (Eds.), *Talk and social interaction* (pp. 86–100). Clevedon: Multilingual Matters.

Kasper, G. (1997). "A" stands for acquisition: A response to Firth and Wagner. *The Modern Language Journal, 81,* 307–312.

Kasper, G. (2002). Conversation analysis as an approach to second language acquisition: Old wine in new bottles? Invited talk, SLATE speaker series, University of Illinois at Urbana-Champaign, March 13, 2002.

Kasper, G. & Ross, S. (2001). Is drinking a hobby, I wonder: Other-initiated repair in language proficiency interviews. Paper presented at the Annual AAAL Conference, St. Louis, MI. March 2001.

Kormos, J. (2001). The role of attention in monitoring second language production. *Language Learning, 50,* 343–384.

Koshik, I. (2002a). Designedly incomplete utterances: A pedagogical practice for eliciting knowledge displays in error correction sequences. *Research on social interaction, 35*(3), 277–309.

Koshik, I. (2002b). A conversation analytic study of yes/no questions which imply reversed polarity assertions. *Journal of Pragmatics, 34,* 1851–1877.

Koshik, I. (2003). Wh questions used as challenges. *Discourse Studies, 5*, 51–77.

Kowal, M. & Swain, M. (1994). Using collaborative production tasks to promote students' language awareness. *Language Awareness, 3,* 73–93.

Krashen, S. D. (1980). The input hypothesis. In J. E. Alatis (Ed.), *Current issues in bilingual education* (pp. 168–180). Washington, DC: Georgetown University Press.

Krashen, S. D. (1981). *Second language acquisition and second language learning.* Oxford: Pergamon.

Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Oxford: Pergamon.

Krashen, S. D. (1985). *The input hypothesis.* London: Longman.

Krashen, S. D. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *Modern Language Journal, 73,* 440–464.

Lerner, G. H. (1995). Turn design and the organization of participation in instructional activities. *Discourse Processes, 19,* 111–131.

Lazaraton, A. (2003). Incidental displays of cultural knowledge in the nonnative-English-speaking teacher's classroom. *TESOL Quarterly*, *37*, 213–245.

Lazaraton, A. (2004). Gesture and speech in the vocabulary explanations of one ESL teacher: A microanalytic inquiry. *Journal of Language Learning, 54*(1), 79–117.

Liddicoat, A. (1997). Interaction, social structure, and second language use: A response to Firth and Wagner. *The Modern Language Journal, 81*, 313–317.

Lightbown, P. M. & Spada, N. (1990). Focus-on-form and corrective feedback in communicative language teaching: Effects on second language learning. *Studies in Second Language Acquisition, 12,* 429–448.

Long, M. H. (1980). Input, interaction and second language acquisition. PhD Dissertation, University of California at Los Angeles.

Long, M. H. (1981). Input, interaction and second language acquisition. In H. Winitz (Ed.), *Native language and foreign language acquisition* (pp. 259–278). *Annals of the New York Academy of Sciences, 379*, 259–278.

Long, M. H. (1983a). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics, 4*, 126–141.

Long, M. H. (1983b). Linguistic and conversational adjustments of non-native speakers. *Studies in Second Language Acquisition, 5,* 177–193.

Long, M. H. (1985a). Input and second language acquisition theory. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 377–393). Rowley, MA: Newbury House.

Long, M. H. (1985b). A role for instruction in second language acquisition: Task-based language training. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 77–99). Clevedon: Multilingual Matters.

Long, M. H. (1989). Task, group, and task-group interactions. *University of Hawai'i Working papers in ESL, 8,* 1–26.

Long, M. H. (1990) The least a second language acquisition theory needs to explain. *TESOL Quarterly, 24*, 649–666.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. B. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in perspective* (pp. 39–51). Amsterdam: John Benjamins.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 414–468). New York, NY: Academic Press.

Long, M. H. (1997). Construct validity in SLA research: A response to Firth and Wagner. *The Modern Language Journal, 81,* 318–323.

Long, M. H. (1998). SLA: Breaking the siege. Plenary address, PacSLRF 3, Tokyo, Japan: Aoyama Gakuin University, March, 1998. *University of Hawai'i Working Papers in ESL, 17*(1), 79–129.

Long, M. H., Adams, L. McLean, M., & Castaños, F. (1976). Doing things with words – Verbal interaction in lockstep and small group classroom situations. In J. F. Fanselow & R. Crymes (Eds.), *On TESOL '76* (pp. 137–153). Washington, DC: TESOL.

Long, M. H. & Crookes, G. (1992) Three approaches to task-based syllabus design. *TESOL Quarterly, 26,* 27–56.

Long, M. H. & Crookes G. (1993) Units of analysis in syllabus design: The case for task. In G. Crookes & S. M. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice* (pp. 9–54). Clevedon: Multilingual Matters.

Long, M. H., Inagaki, S., & Ortega, L. (1998). The role of implicit negative feedback in SLA: Models and recasts in Japanese and Spanish. *The Modern Language Journal, 82,* 357–371.

Long, M. H. & Porter, P. A. (1985). Group work, interlanguage talk and second language acquisition. *TESOL Quarterly, 19*, 207–228.

Long, M. H. & Sato, C. (1983). Classroom foreigner talk discourse: Forms and functions of teachers' questions. In H. W. Seliger & M. H. Long (Eds.), *Classroom-oriented research in second language acquisition* (pp. 268–285). Rowley, MA: Newbury House.

Lyster, R. (1998). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning, 48,* 183–218.

Lyster, R. & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition, 19,* 37–66.

Mackey, A. & Philp, J. (1998). Conversational interaction and second-language development: Recasts, responses and red herrings? *The Modern Language Journal, 82,* 338–356.

Markee, N. (1994). Toward an ethnomethodological respecification of second language acquisition studies. In E. Tarone, S. Gass, & A. Cohen (Eds.), *Research methodology in second language acquisition* (pp. 89–116). Hillsdale, NJ: Lawrence Erlbaum.

Markee, N. (1995). Teachers' answers to students' questions: Problematizing the issue of making meaning. *Issues in Applied Linguistics, 6,* 63–92.

Markee, N. (2000). *Conversation analysis.* Mahwah, NJ: Lawrence Erlbaum.

Markee, N. (2003). Qualitative Research Guidelines (Conversation Analysis). In C. Chapelle & P. Duff (Eds.), Some guidelines for conducting quantitative and qualitative research in TESOL. *TESOL Quarterly, 37*, 169–172.

Markee, N. (2004). Zones of interactional transition. *The Modern Language Journal*, *88,* 583–596.

Markee, N. (2005a). Conversation analysis for second language acquisition. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (pp. 355–374). Mahwah, NJ: Lawrence Erlbaum.

Markee, N. (2005b). The organization of off-task talk in second language classrooms. In K. Richards & P. Seedhouse (Eds.), *Applying conversation analysis* (pp. 197–213). Basingstoke: Palgrave MacMillan.

McGroarty, M. (1998). Constructive and constructivist challenges for applied linguistics. *Language Learning, 48,* 591–622.

McHoul, A. (1978). The organization of turns at formal talk in the classroom. *Language in Society, 7*, 183–213.

McHoul, A. (1990). The organization of repair in classroom talk. *Language in Society, 19,* 349–377.

Mehan, H. (1979). *Learning lessons: social organization in the classroom.* Cambridge, MA: Harvard University Press.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20,* 83–108.

Mori, J. (2002). Task design, plan, and development of talk-in-interaction: An analysis of a small group activity in a Japanese language classroom. *Applied Linguistics, 23*, 323–347.

Muranoi, H. (2000). Focus on form through interaction enhancement: Integrating formal instruction into a communicative task in EFL classrooms. *Language Learning, 50,* 617–673.

Nicholas, H., Lightbown, P., & Spada, N. (2001). Recasts as feedback to language learners. *Language Learning, 51,* 719–758.

Nunan, D. (1993). Task-based syllabus design: Selecting, grading and sequencing tasks. In G. Crookes & S. M. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice* (pp. 55–68). Clevedon: Multilingual Matters.

Ohta, A. (2001a). *Second language acquisition processes in the classroom.* Mahwah, NJ: Lawrence Erlbaum.

Ohta, A. (2001b). Confirmation checks: A conversation analytic reanalysis. Paper presented as part of the colloquium on *Unpacking negotiation: A conversation analytic perspective on L2 interactional competence*, Annual AAAL Conference, St. Louis, MI, March 2001.

Oliver, R. (1995). Negative feedback in child NS-NNS conversation. *Studies in Second Language Acquisition, 17,* 459–481.

Oliver, R. (1998). Negotiation of meaning in child interactions. *Modern Language Journal, 82,* 372–386.

Oliver, R. (2000). Age differences in negotiation and feedback in classroom and pair work. *Language Learning, 50,* 119–151.

Olsher, D. (2001). Negotiating plans of action in small group interaction. Paper presented as part of the colloquium on *Unpacking negotiation: A conversation analytic perspective on L2 interactional competence*, Annual AAAL Conference, St. Louis, MI, March 2001.

Peck, S. (1978). Child-child discourse in second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 383–400). Rowley, MA: Newbury House.

Peck, S. (1980). Language play in child second language acquisition. In D. Larsen-Freeman (Ed.), *Discourse analysis in second language research* (pp. 154–164). Rowley, MA: Newbury House.

Pica, T. (1987). Second language acquisition, social interaction and the classroom. *Applied Linguistics, 8,* 3–21.

Pica, T. (1992). The textual outcomes of native speaker/non-native speaker negotiation. What do they reveal about second language learning? In C. Kramsch & S. McConnell-Ginet (Eds.), *Text in context: Crossdisciplinary perspectives on language study* (pp. 198–237). Lexington, MA: D. C. Heath.

Pica, T. & Doughty, C. (1985). The role of group work in classroom second language acquisition. *Studies in Second Language Acquisition, 7,* 233–248.

Pica, T., Doughty, C., & Young, R. (1986). Making input comprehensible: Do interactional modifications help? *IRAL, 72,* 1–25.

Pica, T., Holliday, A., Lewis, L., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition, 11,* 63–90.

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language instruction. In G. Crookes & S. M. Gass (Eds.), *Tasks and language learning: Integrating theory and practice* (pp. 9–34). Clevedon: Multilingual Matters.

Porter, P. (1986). How learners talk to each other: Input and interaction in task-centered discussions. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 200–222). Rowley, MA: Newbury House.

Robinson, J. D. (1998). Getting down to business: Talk, gaze and body orientation during openings of doctor-communication consultations. *Human Communications Research, 25,* 97–123.

Robinson, P. J. (1995). Attention, memory and the 'noticing' hypothesis. *Language Learning, 45,* 283–331.

Roger, D. & Bull, P. (1988). Introduction. In D. Roger & P. Bull (Eds.), *Conversation* (pp. 21–47). Clevedon: Multilingual Matters.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language, 50,* 696–735.

Salica, C. (1981). Testing a model of corrective feedback. MA thesis, University of California at Los Angeles.

Sato, C. (1986). Conversation and interlanguage development: Rethinking the connection. In R. R. Day (Ed.), *Talking to learn* (pp. 23–45). Rowley, MA: Newbury House.

Sato, C. (1988). Origins of complex syntax in interlanguage development. *Studies in Second Language Acquisition*, *10,* 371–395.

Schegloff, E. A. (1968). Sequencing in conversational openings. *American Anthropologist, 70,* 1075–1095.

Schegloff, E. A. (1979). The relevance of repair to syntax-for-conversation. In T. Givón (Ed.), *Syntax and semantics,* Volume 12: *Discourse and Syntax* (pp. 261–286). New York, NY: Academic Press.

Schegloff, E. A. (1991). Conversation analysis and socially shared cognition. In L. R. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Socially shared cognition* (pp. 150–171). Washington, DC: American Psychological Association.

Schegloff, E. A. (1992). Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology,* 97, 1295–1345.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction, 26*, 99–128.

Schegloff, E. A. (1997a). Third turn repair. In G. R. Guy, C. Feagin, D. Schiffrin, & J. Baugh (Eds.), *Towards a social science of language: papers in honor of William Labov*. Volume 2: *Social interaction and discourse structures* (pp. 31–40). Amsterdam: John Benjamins.

Schegloff, E. A. (1997b). Practices and actions: Boundary cases of other-initiated repair. *Discourse Processes, 23,* 499–545.

Schegloff, E. A. (2000). When "others" initiate repair. *Applied Linguistics, 21,* 205–243.

Schegloff, E. A. & Sacks, H. (1973). Opening up closings. *Semiotica*, *8,* 289–327.

Schegloff, E., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language, 53*, 361–382.

Schegloff, E. A., Koshik, I., Jacoby, S., & Olsher, D. (2002). Conversation analysis and applied linguistics. In M. McGroarty (Ed.), *ARAL, 22, Discourse and dialog*, 3–31.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11,* 17–46.
Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics, 13,* 206–226.
Schmidt, R. (Ed.). (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review, 11,* 11–26.
Schmidt, R. & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner. In R. R. Day (Ed.), *Talking to learn* (pp. 237–326). Rowley, MA: Newbury House.
Schwartz, J. (1980). The negotiation for meaning: Repair in conversations between second language learners of English. In D. Larsen-Freeman (Ed.), *Discourse analysis in second language research* (pp. 138–153). Rowley, MA: Newbury House.
Seedhouse, P. (1997). The case of the missing "No": The relationship between pedagogy and interaction. *Language Learning, 47,* 547–583.
Seedhouse, P. (1999). The relationship between context and the organization of repair in the L2 classrooom. *IRAL, XXXVII,* 59–80.
Sharwood-Smith, M. (1991). Speaking to many minds: On the relevance of different types of language information for the L2 learner. *Second Language Research, 7*, 118–132.
Sharwood-Smith, M. (1993). Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition*, *15,* 165–179.
Shehadeh, A. (1999). Non-native speakers' production of modified comprehensible output and second language learning. *Language Learning, 49,* 627–675.
Sinclair, J. & Coulthard, M. (1975). *Towards an analysis of discourse.* Oxford: OUP.
Skehan, P. & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1,* 185–221.
Spada, N. (1997). Form-focused instruction and second language acquisition: A review of classroom laboratory research. *Language Teacher, 30,* 73–87.
Spada, N. & Lightbown, P. (1993). Instruction and the development of questions in L2 classrooms. *Studies in Second Language Acquisition, 15,* 205–224.
Stivers, T. (2001). Negotiating who presents the problem: Next speaker selection in pediatric encounters. *Journal of Communication, 51,* 252–282.
Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Rowley, MA: Newbury House.
Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford: OUP.
Swain, M. & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics, 16,* 371–391.
Tarone, E. & Liu, G. (1995). Situational context, variation, and second language acquisition. In G. Cook & B. Seidlhoffer (Eds.), *Principles and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 107–124). Oxford: OUP.
Tomasello, M. & Herron, C. (1988). Down the garden path: Inducing and correcting over-generalization errors in the foreign language classroom. *Applied Psycholinguistics, 9,* 237–246.

VanPatten, B. (1988). How juries get hung: Problems with the evidence for a focus on form in teaching. *Language Learning, 38*, 243–260.

Varonis, E. M. & Gass, S. M. (1985a). Miscommunication in native/nonnative conversation. *Language in Society, 14,* 327–343.

Varonis, E. M. & Gass, S. M. (1985b). Non-native/non-native conversations: A model for negotiation of meaning. *Applied Linguistics, 6,* 71–90.

Wagner, J. (1996). Foreign language acquisition through interaction: A critical review of research on conversational adjustments. *Journal of Pragmatics, 23,* 215–235.

White, J. & Lightbown, P. M. (1984). Asking and answering in ESL classrooms. *Canadian Modern Language Review, 40,* 228–244.

White, L. (1987). Against comprehensible input: The input hypothesis and the development of second language competence. *Applied Linguistics, 8,* 95–110.

White, L. (1989). *Universal grammar and second language acquisition.* Amsterdam: John Benjamins.

White, L. (1991). Adverb placement in second language acquisition: Some effects of positive and negative evidence in the classroom. *Second Language Research, 7,* 133–161.

White, L., Spada, N., Lightbown, P., & Ranta, L. (1991). Input enhancement and L2 question formation. *Applied Linguistics, 12,* 416–432.

Willey. B. (2001). Examining a "communications strategy" from a conversation analytic perspective: Eliciting help from native speakers inside and outside of word search sequences. MA thesis, University of Illinois at Urbana-Champaign.

Williams, J. (1999). Learner-generated attention to form. *Language Learning, 49,* 583–625.

Yule, G. & McDonald, D. (1990). Resolving referential conflicts in L2 interaction: The effect of proficiency and interactive role. *Language Learning, 40,* 539–556.

PART II

Discussion

Generalizability

A journey into the nature of empirical research in applied linguistics*

Lyle F. Bachman
UCLA

In this response I discuss generalizability and some issues this raises about research in applied linguistics. I describe generalizability in terms of inferential links from an observation to a report, to an interpretation, to the uses of that interpretation. In considering how to articulate, and support with evidence, these links, three aspects of generalizability need to be addressed: consistency, meaningfulness and consequences.

Five dimensions of research can influence the ways in which generalizability is addressed by different researchers: (1) the researcher, (2) the entity of interest, (3) the context, (4) the observation and report, and (5) the researcher's interpretation. Consideration of these dimensions can help capture the richness and complexity of the varied approaches to empirical research in applied linguistics.

A "research use argument" can guide research design and provide the justification for the interpretations and uses of research. Adopting an epistemology of argumentation would move us away from one that is driven by a preoccupation with our differences toward one that admits a variety of research methodologies. It is clear that the use of multiple approaches in research is both feasible and desirable, in that it expands the methodological tools that are at the researcher's disposal, thereby increasing the likelihood that the insights gained from the research will constitute genuine advances in our knowledge.

Introduction

After having read these very thoughtful and stimulating chapters, I must say that my first reaction was one of being overwhelmed at the enormity of what we applied linguists are presuming to accomplish. When I consider the conceptual

conundrums over which we cogitate and the methodological mazes through which we meander, I'm inclined to advise my students to go into some simpler endeavor, something less complex and relatively straight-forward, like rocket science. After all, launching an electronic explorer on a trajectory to rendezvous with a distant planet in 25 year's time is a piece of cake compared to identifying the specific learning challenges for a given language learner, determining what kinds of language use activities will provide the most effective interactions for him or her, how a teacher can best implement these, and then assessing how much language that learner has learned after a program of instruction.

The questions that the chapter authors were asked to address had to do with generalizability of results, appropriateness of inferences, and dependability in the research process, and the chapters in this volume consider these notions from a variety of research perspectives within applied linguistics. Superficially, one might conclude that these notions are essentially the same across these research approaches, simply playing themselves out differently at the levels of methodology and analysis. However, probing more deeply, the consideration of these questions forces us to delve into the epistemological underpinnings of our various approaches to research; indeed, they go to the very core of our world views as researchers.

There are perhaps two propositions about which we all seem to agree. First, empirical research is aimed at the creation or construction of knowledge. Chapelle (this volume), for example, argues that "every study contributes individual pieces to the store of professional knowledge" (p. 48), while Duff (this volume) states that the "aim of research is to generate new insights and knowledge" (p. 66). Second, there seems to be general agreement that we create this knowledge by observing phenomena. Thus Duff (this volume), states that "research in education and the social sciences has long been concerned with the basis for inferences and conclusions drawn from empirical studies" (p. 65), while Swain (this volume) speaks of drawing "inferences from the data collected" (p. 97). Similarly, with respect to language assessment, McNamara (this volume) states that "it is fair to characterize all such research aiming at providing evidence for or against the legitimacy of inferences from test scores, both qualitative and qualitative, as empiricist" (p. 166). However, despite these points of agreement, the chapters in this volume demonstrate that there are considerable differences in what counts as "knowledge", in how we go about creating it, and in the values and assumptions that underlie our different approaches to conducting empirical research.

The authors of the chapters in this volume address the issues of generalizability from a variety of perspectives, and use the term "generalizability" in a

number of different ways. Most often, the term is used to refer to the inferential link between an observation and an interpretation. However, it is also used to refer to an inference from the observation of a single case to a larger group, and sometimes it refers to the dependability or consistency of the observations themselves. Several of the authors point out the limitations of viewing consistency as a desirable quality of observations, and discuss other qualities, such as trustworthiness, credibility, verisimilitude, authenticity, and comparability, as being equally, if not more, important.

In this response, I discuss what I see as the nature of generalizability, and some of the issues this raises about the nature of research. I would hasten to point out that, as might be expected in a response chapter, there is a some overlap between what I discuss here and the discussions of generalizability in the chapters in this volume. The chapters by Duff, Chapelle and McNamara, in particular, provide excellent overview discussions of many of the issues I address here. Thus, what I discuss in this response chapter is partly a selective discussion of issues they raise. At the same time, I believe that what I present here reflects a slightly different perspective on issues of generalizability from the views presented in these chapters, as well as in the other chapters in this volume.

I first lay out a broad view of generalizability in terms of logical links from observed performance to an observation report, to an interpretation, to the use and consequences of that interpretation. I argue that in addition to considering generalizability as consistency, or reliability, across observations, and as the meaningfulness, or validity, of the interpretations we make from these observations, we also need to consider the uses we make of our research results, and the consequences of this use for various individuals who may be affected by these uses.

I then discuss five dimensions of the research process that I believe determine the ways in which these different aspects of generalizability are addressed by different researchers and in different research studies. I argue that these dimensions need to be accounted for or at least considered when we address the generalizability of our research results. I also argue that these dimensions can help capture the richness and complexity of the varied approaches to empirical research in applied linguistics, and provide a way of getting beyond the "quantitative-qualitative, positivist-constructivist" division that has long been recognized as an oversimplification of the diversity of research approaches in applied linguistics. I then describe a "research use argument" and propose this as a basis both for guiding the conceptualization or design of research studies in applied linguistics, and for providing the justification for interpretations

and uses of research results. Finally, I offer some observations on the viability of combining multiple approaches to research in applied linguistics.

What is the nature of "generalizability"?

In most empirical research in applied linguistics, we are interested in linking observations of phenomena to interpretations. That is, we want to attach meaning to our observations. In our field, these observations are typically of language use, or the performance of language users. In addition to attaching meaning to an observation of performance, in much applied linguistics research this meaning, or interpretation, is used, sometimes by the researcher, sometimes by others, to make a generalization or a decision, or to take an action that goes beyond the observation and its particular setting. Thus, in much of our research in applied linguistics, we need to make a series of *inferences* that link performance to use. The inferential links between a performance and our use of that performance are illustrated in Figure 1, after Bachman (2004a).

The **observed performance** in the bottom box in this figure is an instance of the observable phenomenon upon which we are focusing our research. This performance might be, for example, a conversation, a "collaborative dialogue", a group interaction among members of a speech community (e.g., family, classroom, applied linguists at conference presentation), a grammaticality judgment, or performance on an assessment task. The **observation report** in the next box up, is the initial extraction, or filtering, by the researcher, of the performance that is observed. This observation report might consist of, for example, a score, a verbal description, a transcription, a picture or a video clip. This observation report is *inferred* from the observation. This is because, in order to arrive at the observation report, the researcher makes decisions about what to observe and how to record it. These decisions "filter" both the performance to be observed out of all the possible performances that might be observed, and the observations themselves, since even the most detailed observation report will necessarily reflect the selective perceptions of the researcher. Other researchers might make different decisions, filter the performance differently, and hence arrive at, or infer, different observation reports.

The researcher arrives at the observation report by deciding what to observe and by deciding to use observation and analytic procedures that are consistent with and appropriate for the particular methodology that is employed. These decisions about what and how to observe, and what and how to transcribe or report will be informed by the researcher's own methodolog-

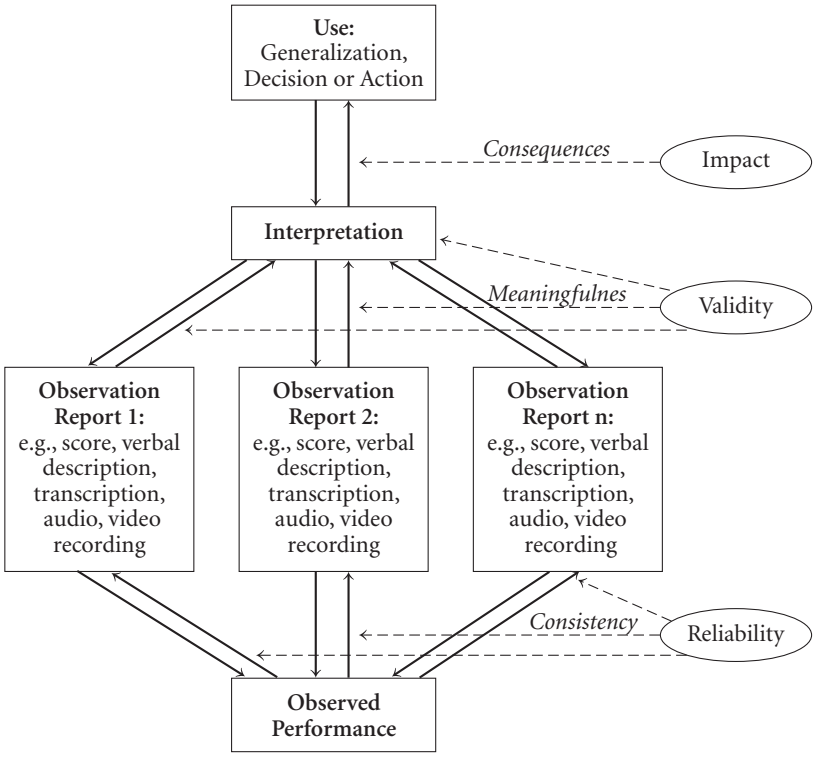

Figure 1. Inferential links from an observation to use (after Bachman 2004a: 725)

ical and theoretical perspectives, or biases (Ochs 1979). With a conversation analysis, this methodology will include the way in which the particular conversation is selected and the medium with which it is recorded, as well as the procedures employed to transcribe the details (e.g., words, non-verbal sounds, pauses) of the conversation. With an ethnography, this will consist of the reason for selecting a particular group for observation, the roles of the observer and the other participants, the medium with which the interaction is recorded, and the procedures for transcribing or otherwise making field notes about the verbal and non-verbal information in the interactions. For a language assessment, the methodology will consist of the how the assessment developer has defined the ability or attribute to be assessed, the kinds of assessment tasks that are presented to the test takers, the administrative procedures, and the scoring method used. The rigor with which the observation and analytic process are described and followed by the researcher provides the primary justification for the link between the performance and the observation report.

If we assume that a given observation report is but one instance of a number of possible observation reports based on the performance, then we might be interested to know whether this particular report might generalize to a theoretically infinite universe of all such reports, as illustrated in the three boxes in the middle of the figure. Other similar reports might be transcriptions of a conversation by different researchers, verbal descriptions of an interaction by different observers or participants in a speech event, or scores obtained from different forms of a test aimed at measuring the same ability. Generalizability in this sense is related to **consistency** of observation reports, which is often associated with the quality of **reliability**.

The next box up in the chain is the **interpretation**, or the meaning that we want to infer from our observation report. This interpretation might consist of the recognition of a recurring pattern in the observations, to which the researcher might assign a label, such as "adjacency pair", "knowledge of L2 vocabulary", or "the ability to use the existential 'there' appropriately in English". Inferring meaning from an observation report involves a broader type of generalization: from very concrete verbal descriptions or narratives, visible images, or numbers, to an inferred meaning, which will generally be an abstract concept. Generalizability in this sense pertains to the **meaningfulness**, or **validity**, of the interpretation.

The top box is the **use** the researcher or consumers of the research results make of the interpretation. The inference from interpretation to use typically consists of a generalization to a larger group or to other situations, or a decision, or an action or implication for action, that the researcher or consumers of the research results make on the basis of the interpretation. Generalizability in this sense pertains to the **consequences**, or **impact**, of the use of the interpretation.

In summary, I would argue that in applied linguistics research, the links between a performance, an observation report, an interpretation, and a use, consist essentially of a chain of logical inferences, or generalizations, from one level to the next. I would also argue that each inference in the chain must be supported both by logical arguments and empirical evidence in support of those arguments.[1] I further argue that the inferential and evidentiary burden increases as the researcher's purpose moves from that of describing the performance in the observation report, to interpreting that report, to using this interpretation in some way.

If the researcher's objective is to provide a careful and detailed *description* of the phenomenon, she will need to consider the consistency, or reliability of that description, with respect to how it might differ from those of other re-

searchers. Two conversation analysts, for example, might differ in how they transcribe particular details of a conversation, and different participants in a speech event might provide different reports of that event. Similarly, scores for an individual from two different elicitations of grammaticality judgments, or from two different forms of a language test intended to assess the same aspect of language ability, might differ. The extent to which such inconsistencies are considered problematic, and how they are resolved, differ across the different research approaches that are commonly used in applied linguistics. In some approaches, inconsistencies may be regarded as beneficial, in that they provide differing perspectives. The building up of multiple observations into a thick description in ethnography, for example, enables the researcher to report both commonalities and differences in the observations of different participants. In other approaches, such as experimental or quasi-experimental designs, or language assessment, lack of reliability is generally considered to be a problem, and various statistical procedures have been developed to estimate this and to minimize it.

If the researcher's purpose is to generalize beyond a particular observation, to *interpret* the observation report in a way that gives it meaning within the framework of an existing body of research or a theory, she needs to consider both the consistency of the observation report and the meaningfulness of her interpretation. She will need to articulate a logical argument that she can, indeed, generalize from the observation report (or from multiple observation reports) to the intended interpretation, and then provide evidence to support this argument. The links described above provide the scaffolding she needs for developing a logical argument that the meanings she attaches to her observations are valid. The links also provide a guide for identifying the kinds of evidence she needs to gather in support of this argument. As with consistency, the way the researcher provides convincing support for the meaningfulness of interpretations varies from one research approach to another. In some approaches, this consists essentially of achieving a well-grounded interpretation based on multiple sources of information, while in others, it may also include the statistical analyses of quantitative (numerical) data.

Finally, if the purpose of the research is to provide a basis for a *decision* or an *action*, or if it is highly probable that the research results may have implications for actions that might be taken by the consumers of the research, then the researcher needs to consider the consequences of the decision or action. She, or the consumers of the research, will need to articulate a logical argument that the interpretation of the observation report is relevant to the intended use, that it is actually useful for the intended use, that the consequences of the

intended use are beneficial, and that the probability of potential unintended adverse consequences is minimal.

In considering how to articulate, and support with evidence, the logical links between a given observation, an observation report, an interpretation of the observation report, and a use of that interpretation, there are three issues that need to be addressed: consistency, meaningfulness and consequences.

Consistency of observations and reports

Consistency has to do with the extent to which any given observation report provides essentially the same information, or generalizes, across different aspects, or facets, of the observation and reporting procedure (e.g., instances, events, observers, ethnographers, categorizers, analysts, raters). When considering the consistency of research results, we need to question both the desirability of consistency and the level of consistency that will be acceptable. Is consistency of observation reports necessarily desirable? It is well-known that reliability is an essential quality of measurements. If a score is not a consistent measure of anything, how can it be a measure of something? However, as Swain (1993) has pointed out, concerns with *internal* consistency may limit the capacity of measures for providing information about the construct of interest.[2] In the literature on writing assessment, some researchers view different "readings" of a composition by different raters as a positive outcome (Weigle 2002). Likewise, the "rich description" of ethnography is built up by bringing together multiple perspectives of a given event, all of which are likely to vary in some details, as well as in their emphasis or stance. So whether consistency is viewed as desirable or necessary will reflect the researcher's approach and perspective, as well as her purpose. If the researcher's purpose is to interpret different observation reports, or scores, as indicators or descriptions of essentially the same construct, as is the case with language assessment, then consistency among reports is essential. If, however, the purpose is primarily to describe the phenomenon in all its richness, as in conversational analysis and ethnography, then variations in reports may be viewed as adding to and enhancing the description of the phenomenon.[3]

If consistency is considered desirable or necessary, then another question that we need to address is, "What level of consistency is acceptable?" How consistent, and in what ways, do the observations, or data collection procedures need to be? How consistent, and in what ways, do test scores, transcriptions, ethnographies, analyses, or reports of different observers need to be, in order

to be reliable, trustworthy, or credible? The answer to this question will depend, of course, on the importance and types of decisions or actions to be taken. If one is making potentially life affecting decisions about large numbers of people, as in high-stakes testing, we need to have highly reliable measures. However, if we are testing a theoretical hypothesis in research, how consistent do our reports need to be? What if were are primarily interested in providing a description of a particular phenomenon? How trustworthy does this description need to be? The way we answer these questions will determine how we view and operationalize generalizability as consistency in our research, whether this is in the form of a statistical estimate of reliability, or a consensus of researcher observations.

Meaningfulness of interpretations

When a researcher interprets her observations, she is, in effect, attaching meaning to these with respect to some domain of prior research or experience, or a theoretical position, in the field. Thus, the research may enrich her understanding of general processes involved in interaction, or of co-constructing discourse, or it may lead to a new generalization about these. But while the researcher may feel very confident about the meaningfulness of her interpretation, how do other researchers evaluate her claim of meaningfulness? Duff (this volume) provides an excellent discussion of meaningfulness, or validity, in quantitative and qualitative research, and I will thus provide only a brief summary of these issues here.

Meaningfulness in quantitative research

Within quantitative approaches to research, meaningfulness has been conceptualized from two perspectives: research design and measurement. In a classic article, Campbell and Stanley (1963) describe two types of meaningfulness, or validity, in experimental and quasi-experimental designs: external validity and internal validity. **External validity** is what might be called generalizability or extrapolation from the results of a study based on a sample, to a population or to other similar settings. External validity thus has to do with the extent to which the research results generalize beyond the study itself. **Internal validity**, on the other hand, is the extent to which the causal connection between treatment and outcome that is inferred from the results is supported by features of the research design itself (e.g., randomization, distinctiveness of treatments,

non-interaction between treatments). In measurement, **validity** refers to the meaningfulness and appropriateness of interpretations (Messick 1989). In the current "standard" paradigm for measurement, validity is seen as a unitary concept, and evidence in support of the validity or meaningfulness of interpretations can be collected in a variety of ways (e.g., content coverage, concurrent relatedness, predictive utility).[4] Much current research in language testing is conducted within this paradigm, but as McNamara (this volume) points out in his chapter, this view of validity is quite narrow, in that it is based on essentially quantitative, psychometric methods. However, recent argument-based approaches to conceptualizing validity in educational measurement have taken a much broader view of validity, and provide for the inclusion of multiple types of evidence, both qualitative and quantitative, to support a validity argument (e.g., Kane 2001; e.g., Kane 2004; Kane et al. 1999; Mislevy 2003; Mislevy et al. 2003). Bachman (2005) has extended this argument-based approach so as to link not only observations with interpretations, but also to provide a logical means for linking interpretations to intended uses – the decisions or actions that may be taken on the basis of interpretations – and the consequences of these decisions or actions.

Meaningfulness in qualitative research

Within the approaches to research that are broadly referred to as "qualitative", meaningfulness, or validity, is also of concern, although the way it is conceptualized and implemented in practice differs, understandably, from that of quantitative research design and measurement. As with quantitative research, validity within the qualitative research tradition has been conceived from several different perspectives.[5]

Kirk and Miller (1999) define the issue of validity in qualitative research as "a question of whether the researcher sees what he or she thinks he or she sees" (p. 21). In a similar vein, Lynch (1996), citing Maxwell (1992), defines validity as "the correspondence between the researcher's 'account' of some phenomenon and their 'reality' (which may be the participant's constructions of the phenomena)" (p. 55). Lynch discusses several different types of validity, as defined by Maxwell: descriptive validity (the factual accuracy of the research account), interpretive validity (how accurately the account describes what the phenomenon means to the participants – excluding the researcher), theoretical validity (how well the account explains the phenomenon), generalizability (internal generalizations within the group being studied and external generalization to other groups) and evaluative validity (how accurately the research

account assigns value judgments to the phenomenon) (Lynch 1996: 55). Drawing on the research in naturalistic inquiry (Guba & Lincoln 1982) Lynch also discusses validity in terms of "trustworthiness criteria": credibility, transferability, dependability and confirmability (1996: 56). Lynch describes a variety of techniques for assessing and assuring validity in qualitative/naturalistic research (e.g., prolonged engagement, persistent observation, negative case analysis, thick description, dependability audit and confirmability audit) (p. 57). Lynch also discusses the use of triangulation, which he describes as "the gathering and reconciling of data from several sources and/or from different data-gathering techniques" (Lynch 1996: 59). Citing the limitations of the navigational metaphor of triangulation, Lynch discusses another metaphor for this approach, attributed to Mathison (1988), that of detective work. In this metaphor, the researcher sifts through the data collected from multiple sources, looking not only for agreement, but also searching for examples of disagreement. In this way, triangulation is not a confirmation of the meaningfulness through the collection of research results that converge, but is rather "the search for an explanation of contradictory evidence" (Lynch 1996: 61).

In summary, meaningfulness, or validity, is of paramount interest and concern in both quantitative and qualitative/naturalistic approaches to research. For both research approaches, the essential question is the extent to which the researcher's interpretation of, or the meaning he attaches to the phenomenon that is observed, can be justified, in terms of the evidence collected in the research study. In evaluating the meaningfulness of his interpretations, the researcher is, in effect, constructing a logical argument that his interpretation is justified, and collecting evidence that is relevant to supporting or justifying this interpretation. While the logic and structure of the validity argument may well be consistent irrespective of the research approach used, the types of evidence collected and the methods for collecting these will vary considerably across different research approaches, and will be highly specific to any given research study.

Consequences of decisions or actions for stakeholders

The research that we conduct as applied linguists may lead to a decision or action, based on the meaning that is attached to the observation. These decisions or actions will have an impact on, or consequences for, different groups of stakeholders.[6] If, for example, the interpretation of our observation report leads to placing in question a particular hypothesis or proposition that might

be part of a theory or widely-held view in the field, it may result in a reformulation of the theory or view, or in its rejection by the field. This, in turn, may affect the research that is subsequently conducted, and that gets published by other researchers, or the ways in which the results of other research are interpreted. In much of our research as applied linguistics, however, the intended use of the research is to find solutions to real world problems. This applied purpose is explicit in Larsen-Freeman's intended uses of grammatical explanations that are derived from data, for example, "to inform the identification of the language acquisition/learning challenge of language students" (Larsen-Freeman this volume, p. 1). It is also clearly implicit in Duff's statement, "while the research may speak to issues of how they [language learners] would engage in one particular type of task, it does not shed light on how curriculum can be developed linking such tasks in meaningful, educationally sound ways" (Duff this volume, p. 6).

Thus, in much applied linguistics research, the use we make of an interpretation takes the form of an action or an implication for action in the real world. These actions will have consequences for individuals beyond those whose performance was part of our research. For example, language educators may use our interpretations of classroom interactions, or of performance on grammaticality judgments, to implement changes in language curricula and pedagogical practice. Or, a language program administrator might interpret a score on a test as an indicator of high language aptitude and on this basis place students with high test scores into an intensive language course. Or, test users may use our interpretations of performance on a language test to make decisions about certifying individuals' professional qualifications, about issuing them a visa, or about hiring them. When our interpretations are used to make decisions or actions beyond the research itself, we need to extend our notion of generalizability to include the consequences of the way we use our results. That is, we need to consider the **impact** of our research on those who will be affected our decisions or implications. This is of particular importance when the lives of other individuals, such as language learners, teachers or students, or specific groups of language users, may be affected by the uses that are made of our research results.

The researcher's responsibility for consequences

If the results of our research do, indeed, have consequences, or impact, beyond the research itself, then it would be reasonable to consider the question of the extent or limit of the researcher's responsibility for the way research re-

sults are used. It is in the field of language assessment, where language testers are most often held accountable for the ways in which the results of their tests are used, that the issues of responsibility for consequences has been discussed most extensively. In the past decade or so there has been an expanding discussion of ethics and consequences in the language testing literature, and a full review of this is neither feasible nor appropriate here.[7] Rather, I will focus on the discussions of responsibility that seem to be most relevant and generalizable to research in applied linguistics. Davies (1997a) and Hamp-Lyons (1997a) both discuss the language tester's (researcher's) responsibility for the way the results of a given test (research study) are used. Hamp-Lyons (1997b) argues that language testers have responsibility "for all those consequences which we are aware of" (p. 302), leaving unclear what responsibility the language tester has for making herself aware of the possible consequences of test use. Davies (1997a) similarly avoids the issue of the tester's responsibility for anticipating possible misuses of tests when he states that language testers should be willing to be held responsible for "limited and predictable social consequences we can take account of and regard ourselves responsible for" (p. 336). But what about possible unpredictable consequences? Bachman (1990, 2004a) takes a more proactive stance toward the language tester's responsibility for anticipating unintended consequences. Bachman (1990), argues that the language test developer should "list the potential consequences, both positive and negative, of using a particular test" and then "rank these [potential consequences] in terms of the desirability or undesirability of their occurring" (p. 283). Bachman (2004a) takes this a step further, arguing that specific statements about possible unintended consequences need to be articulated as part of an assessment use argument by the test developer. These discussions of consequences and the responsibility of the language test developer are, I believe, directly relevant to other kinds of research in applied linguistics.

If we accept this line of reasoning, then researchers in applied linguistics should articulate, as part of their research design, an argument supporting the potential uses of their research results, in which they indicate what they believe would be appropriate uses and beneficial consequences of their research results, and what they would consider inappropriate, or would have negative consequences for different groups of stakeholders. It is currently quite common for funding agencies to require researchers to provide a statement of the possible impact of their research on various individuals, groups, or institutions. What is not so common, is asking researchers to anticipate what the possible negative impacts of their research might be.

Dimensions of applied linguistics research

I have described generalizability as comprising consistency, meaningfulness and consequences, and have argued that this concept applies to all empirical research in applied linguistics. However, the way in which generalizability is considered will vary from one empirical study to another, and from one researcher to another. In order to better understand the ways in which issues of generalizability are addressed in any given empirical study, I believe it is useful to discuss such research in terms of five dimensions. I would argue that the way in which a particular study or researcher deals with generalizability is a function of where the research and researcher is situated with reference to these dimensions in research. The five dimensions that I will discuss are: (1) the observer/researcher, (2) the entity of interest, (3) the context, (4) the observation and report, and (5) the interpretation. These dimensions are summarized in the Table 1.

Table 1. Dimensions of research in applied linguistics

I. *The observer/researcher*
   - A. What is the researcher's view of the world and perspective on knowledge?
   - B. What is the researcher's purpose for conducting the research?
   - C. Who are the researcher's intended audiences?

II. *The entity of interest: the "construct"*
   - A. What are constructs and where do they reside?
   - B. Where do constructs come from?
   - C. Underspecification and grain-size

III. *The nature and role of "context"*
   - A. What is the relationship between the context and the entity/construct?
   - B. What is the range or scope of context?

IV. *The observation and report*
   - A. What counts as an "observation"?
   - B. What is the unit of analysis?
   - C. How is the observation reported?

V. *The researcher's interpretation of the observation*
   - A. What is the researcher's ontological stance towards the observation?
   - B. What is the relationship between observations and interpretations?

The observer/researcher

The researcher, as observer, will bring to the research enterprise, either explicitly or implicitly, a set of philosophical assumptions about the nature of the world (ontology), and about the nature of our knowledge of that world and about how we acquire that knowledge (epistemology). The researcher will also have a purpose for conducting the research, and one or more audiences to whom she intends to report her results.

What is the researcher's view of the world and perspective on knowledge?
Researchers' views of the nature of the entities that exist or that may exist differ in terms of how they see the relationship between the mind and matter – what has been called the "mind-body problem" (see citations below for monism and dualism). This problem can be articulated as a question: "How is it possible for two entities that appear to be distinct – one apparently physical and the other apparently not – to interact, as our minds seem to do with our bodies?" Stated more broadly, in terms of empirical research, the question becomes, "How is it possible for two distinct entities, the mind of the observer and the object of the observation, to interact, as researchers seem to do with the phenomena they observe?" The way the researcher answers this question leads to two very different ontologies: monism and dualism. **Monism** holds that mind and matter are essentially the same. This view has been associated with philosophers such as Parmenides, Heraclitus, Democritus, the Stoics, Spinoza, Berkeley, Hume, and Hegel, and has been articulated more recently in the work of Churchland (1996), Damasio (1994) and Dennett (1991). The opposing view is that of **dualism**, which holds that mind and matter are fundamentally different. This view has been associated with philosophers such as Plato, Kant, and Descartes, and has been articulated more recently in the work of Chalmers (1996) and Hart (1988).[8]

The ontological view that the researcher adopts has clear implications for epistemology, as well. One implication of a monist view for research is that there is no "objective reality" that is distinct from the researcher, and that can form the basis for our understanding of the world. Thus, the researcher is necessarily a part of the phenomena he wants to investigate. Since each observer may experience his own reality, a corollary of this view is that the researcher accepts the possibility of multiple realities, all of which are equally "true". A dualist ontological view, on the other hand, leads to the epistemological position that the researcher is separate from an objective reality, and that this reality can be discovered by the researcher through the process of observation.

Another distinction that is often discussed in research is that between "etic" and "emic" perspectives. An **emic** perspective is the "insider's" perspective, associated with a participant-oriented approach to knowledge construction.[9] An **etic** perspective, on the other hand, is the "outsider's" perspective, and is typically associated with a researcher-oriented approach to knowledge construction. This distinction has implications for epistemology, if we extend it to claims about knowledge. Conceived in this way, "emic knowledge" represents claims or interpretations expressed in terms that are considered meaningful and appropriate by the members of the particular group that is being observed. The researcher's attention to the emic perspective on knowledge is illustrated, I believe, by Markee's chapter in this collection, when he states explicitly that "the warrant for any analytic claims that are made about how ordinary conversation and institutional talk are organized must be located in the local context of participants' talk, and explicated in terms of what members understand each other to be doing at the time that they are doing it" (p. 12). "Etic knowledge", on the other hand, is the "outsider's" perspective, and represents claims or interpretations expressed from the perspective of accepted practice and terminology in a particular research community. Generalizability of emic knowledge depends largely on consensus among the participants in the speech event, who generally seek to agree that the interpretation is consonant with their shared understanding. Generalizability of etic knowledge, on the other hand, is supported through logical argumentation and the collection of supporting evidence, ideally from multiple sources.

Both perspectives, it seems to me, are illustrated in Swain's chapter. The emic perspective is evident in her use of stimulated verbal protocols to elicit participants' own perceptions of reformulations of stories they have written. The etic perspective, on the other hand, seems to inform her interpretations of these verbalizations as instances of language use – speaking, in this case – mediating language learning. Similarly, both perspectives are implicit in Duff's statement that qualitative research in applied linguistics "seeks to produce an in-depth exploration… in some cases, of participants' and researchers' own positionality and perceptions with respect to the phenomenon" (p. 8).

What is the researcher's purpose for conducting the research?

Research is conducted for a range of purposes. One purpose is to *describe* the phenomenon so as to expand our knowledge or understanding of it. Duff, for example, drawing on Stake's work, says that the purpose of an intrinsic case study is "not to come to understand some abstract construct or generic phenomenon… The purpose is not theory-building" (p. 26). Rather, the in-

trinsic case study is of interest in its own right. Another purpose is to *induce* one or more general statements that might provide a basis for linking the phenomenon observed to other, similar phenomena, or to the phenomenon observed in other contexts, or in other individuals or groups. Here, the researcher goes beyond a description of the phenomenon itself, and attempts to attach meaning to it by generalizing to other individuals or groups of individuals, or to other contexts. An example of this inductive purpose, I believe, can be seen in Duff's depiction of the goal of applied linguistics research: "Rather than understand the phenomenon in terms of its components parts, the goal is to understand the whole as the sum or interaction of the parts" (Duff this volume, p. 8). Another purpose of research is to *explain* the phenomenon, or predict its occurrence or operation in other contexts. This generalization may or may not be related to a particular theory about the phenomenon, but might provide a basis for theory-building. When explanation or prediction is the researcher's purpose, the phenomena to be observed is typically identified on the basis of inductive generalizations from previous research or experience. Another purpose is to use the interpretation to *inform a theory* about the nature of the phenomenon. Chapelle (this volume), for example, discusses linking scores on a vocabulary test to "a broader theoretical construct framework of vocabulary knowledge" (p. 9). In some research, this linkage may entail decisions about the extent to which the theory itself is either supported or placed into question. As McNamara (this volume) points out, "investigation of the relationship between test construct [theory] and test performance cuts both ways: test data… can also be used to question the constructs on which our inferences have been based" (p. 8). If *theory falsification* is the researcher's purpose, the specific phenomenon to be observed will be selected on the basis of the kinds of hypotheses that can be deduced from the theory.

In addition to these various specific purposes, the researcher may have an overarching goal in conducting the research. One goal, which is typically associated with the publication of research results in scholarly journals, is to share the research results with the community of scholars. Since research is never a neutral, objective undertaking, an implicit goal in much published research is to convince other researchers that the findings of the research provide important insights into the phenomenon. Another goal that is typically associated with the "applied" aspect of our discipline, is to use the research results to influence decisions or actions in the "real world" about people or programs. This was discussed above in connection with consequences of research.

Who are the researcher's intended audiences?

When we report the results of our research, the way we report it will depend, to a large extent, on the audience with whom we want to communicate. One audience consists of other members of a particular community of researchers who have been conducting research in the same area as the study we wish to report. Published reports or articles for this audience are likely to be the most focused, and include the most detailed information about the design and procedures followed in the study, the kinds of observations made and reported, and how these were analyzed and interpreted. Articles for this audience are typically published in journals that focus on a specific area of applied linguistics, such as language assessment (e.g., *Language Testing*, *Language Assessment Quarterly*), second language acquisition (e.g., *Language Learning, Studies in Second Language Acquisition*, *Second Language Research*), or discourse analysis (e.g., *Research on Language and Social Interaction*, *Discourse Processes, Journal of Pragmatics*). Another audience consists of the members of a larger community of scholar/researchers. Articles for this audience may address the broader implications of the research results, and are typically published in journals that publish research from a range of areas within a field (e.g., *Applied Linguistics*, *International Review of Applied Linguistics in Language Teaching*, *Issues in Applied Linguistics, TESOL Quarterly*).

In addition to the various research communities with whom we share our research results, there is another, more public, audience, who might be referred to as the "consumers" of our research. These include people and agencies that are in power to make decisions in the real world on the basis of our research. This audience includes individuals who set educational policy at various levels, as well as practitioners; both of these groups may want to use research results to inform policy or practice. In some cases, research results may inform public opinion and influence actions taken by the society at large.

Finally, the results of our research, in the form of publications, research reports and grant applications, are evaluated by committees that consider our research "track record". This audience includes agencies and committees that review grant proposals, as well as committees that evaluate our work with respect to granting tenure and promotion. Thus, any given research report that is published is likely to be read by more than one of these audiences.

The entity of interest: The "construct"

When a researcher observes some phenomenon in the real world, he generally does this because he wants to describe, induce or explain something on the

basis of this observation. That something is what can be called a "construct". When Swain (this volume), for example, asks the question, "What do verbal protocols represent?", she is implicitly looking for an explanation of the phenomenon captured in her verbal protocols in terms of one or more constructs. Constructs are useful in research because they enable the researcher to distinguish the entities that he describes, induces or explains from other entities. I would argue that all research in applied linguistics entails the definition of one or more constructs, even if this is implicit. For example, if a researcher wants to describe a conversation, he needs to start with a definition of what distinguishes this from other types of oral language use, such as an interview, a lecture, or a TV news report. He may also decide to focus only on face-to-face conversations, as opposed to, say, phone conversations, and this, I would argue, also implies an operational definition of the construct, "conversation". Or, if a researcher wants to measure an individual's knowledge of vocabulary, he needs to define this construct in a way that will distinguish it from, say, knowledge of syntax and cohesion. When we define and specify what it is we intend to describe, induce or explain, we are, in effect, constructing a definition of this for the specific purpose of conducting research. And when we construct such a definition, we create a "construct". **Constructs** then, are the entities that we want to describe, induce, or explain, on the basis of our observations of phenomena. The meanings, or interpretations that we infer from our observation reports will be stated in terms of these constructs. As will be seen below, the entities that researchers define as constructs can include the traits, or attributes of the language users, features of the contexts, or interactions between language users and contexts. I would note that Chapelle (1998, this volume) and McNamara (this volume) provide excellent discussions, from the perspective of language assessment, of the role of constructs in research, and of different approaches to defining them.

What are constructs and where do they reside?

The nature of constructs, and where these are believed to "reside" are two issues that are controversial in many areas of applied linguistics. From the perspective of research methodology, the way in which the researcher views these issues is one dimension that characterizes her research approach, and will inevitably be related to her view of the world, as discussed above. Thus, from a monist ontology, the researcher would view the construct as part of a reality within which she is included, while from a dualist perspective, the researcher would view the construct as part of an objective reality that is essentially independent of herself, as researcher. Drawing on Messick (1981), and working within

an implied dualist ontology, Chapelle (1998) has described three alternate approaches to defining constructs, in terms of how researchers interpret consistencies across different observations of phenomena. In a *trait* perspective, consistencies across observations are attributed to characteristics of language users, typically as knowledge, ability, attitudes, feelings, or cognitive processes. In a *behaviorist* perspective, consistencies in observations are attributed to characteristics of the context. Finally, in an *interactionalist* perspective, consistencies across observations are interpreted as reflecting the interactions between traits and contexts.

Trait perspective. From a trait perspective, the researcher would view the construct as a characteristic of the individual language user that can be inferred from consistencies in performance across different observations. From this perspective, the construct "resides" in the individual language user. Conversational turns, adjacency pairs, and repairs in conversation analysis, for example, could be viewed from a trait perspective in terms of the capacity that individual language users have that enables them to perform these aspects of conversational interaction. This is the perspective that Bachman and Palmer (1996) take, for example, when they include "knowledge of conversational organization" as an area of language knowledge. Similarly, Bachman and Palmer view constructs such as morphosyntax, the semantics and pragmatics of a grammatical structure, or patterned use of grammatical features in texts (Larsen-Freeman this volume) as areas of language knowledge that reside in the language users.

Behaviorist perspective. From a behaviorist perspective, the researcher would *identify* the construct with the consistencies in the phenomena that are observed, in which case the construct comprises the characteristics of the phenomena, or could be said to "reside" in the phenomena themselves. Conversational turns, adjacency pairs, and repairs in conversation analysis, for example, could be viewed from a behaviorist perspective essentially as consistencies in the phenomena that the conversational analyst observes. Thus, when Markee (this volume) points out that the construct, "conversational repair" occurs more or less frequently, depending, for example, on the type of task and the types of interactants, it suggests that this construct is a characteristic of conversations, which would imply a behaviorist perspective. Similarly, morphosyntax, the semantics and pragmatics of a grammatical structure, or patterned use of grammatical features in texts (Larsen-Freeman this volume) could be viewed from a behaviorist perspective as constructs that reside in the language *use*

that is observed by the researcher, through introspection, direct observation, or through the analysis of language corpora.

Interactionalist perspective. From an interactionalist perspective, conversational turns, adjacency pairs, morphosyntax, the semantics and pragmatics of a grammatical structure, and so forth, could be seen as constructs that are the products not solely of either the interactants or of the context, but of the interaction between language users and the context. These constructs might be induced through generalization from the observation and analysis of a wide range of conversational or textual discourse phenomena, and that can thus be expected to occur in conversations and other discourse in general. Thus, constructs such as these can also be viewed from an interactionalist perspective as abstract constructs that the conversation or grammar analyst induces from the observations of language use.

Any particular researcher may, either implicitly or explicitly, adopt one of these three perspectives, which will then inform the way he interprets his observations. However, it is important to also understand that different researchers may view essentially the same phenomena from different perspectives and thus interpret them in different ways. Thus, while Bachman and Palmer (1996) may make an inference about the language ability of language users from consistencies in observations of conversational repairs in an oral interview or a group discussion task, it would appear that most conversational analysts would view these consistencies as aspects of the conversations in which they occur. Similarly, it would appear that Larsen-Freeman is interpreting the phenomena she observes in language use as qualities of the texts or discourses in which they occur, rather than as areas of knowledge that language users have.

A variety of constructs, in addition to those mentioned above, are discussed in the chapters in this volume. Coming largely from the field of language assessment, Deville and Chalhoub-Deville, Chapelle, and McNamara provide differing perspectives on how the construct "language ability/proficiency", or some aspect of this, might be defined. Deville and Chalhoub-Deville first describe the "standard" definition of language ability as a fairly stable ability within an individual. Defining constructs this way is essentially the trait perspective on construct definition. However, Deville and Chalhoub-Deville go on to point out the person by task/situation interactions that are typically found in research in measurement, and propose a different, interactionalist, construct, "ability-in-language user-in-context". This construct "does not reside in the head of the test taker, but is bound inextricably to the interaction of the person and the task/context" (p. 14). Chapelle (this volume) discusses several different

dimensions, or traits, that have been proposed for the construct, "vocabulary knowledge", e.g., vocabulary size, organization of L2 lexicon, knowledge of morphological characteristics, and vocabulary strategies, any or all of which might be included in a broad theory of vocabulary knowledge. She also alludes to a construct, "vocabulary ability", which would include both vocabulary knowledge and "the processes for putting the knowledge to use" (Note 1), and which might be considered an interactionalist definition of this construct. McNamara provides a much more general discussion of the relationships among constructs, a "criterion" domain, and test performance. Drawing on different theoretical perspectives on the relationship between cognition and language, Swain (this volume) interprets the phenomena she observes – language learners' verbal protocols of reformulating a story they have written in the target language – in terms of a construct, "mental forms of activity" (p. 13).

Where do constructs come from?

Researchers arrive at construct definitions in a variety of ways. As I've argued above, constructs are always implied by what, where, when and how the observer chooses to observe – by where the researcher chooses to point her camera, and what kind of camera she chooses to use, so to speak. More explicitly, constructs may be induced by generalizing from experience or observation. That is, constructs can be derived from the data. Markee's discussion of the construct, "conversational repair" is, I would argue, and example of a construct that is induced from observation. Another source of constructs is theory. Thus, McNamara (this volume) asserts, at the beginning of his chapter, that in language assessment, "reasoning about the construct… draws on theories of language knowledge and language performance, which are essentially cognitive" (p. 6). In the latter part of his chapter he argues that language testers need to consider the possibility of expanding the ways in which constructs are defined to include both the socio-interactional aspects of language ability, as well as the political ends which language assessments are used to support.

Swain's (this volume) chapter provides an enlightening example of the role of theory in defining constructs, and how the results of empirical observation can serve to clarify theory. Swain traces the implications of two differing theories of mind – cognitivist and sociocultural – for how one would interpret the results of verbal protocols of learners' reformulations of a story, as reported in a collaborative dialogue. What I find most insightful, in terms of generalizability, about Swain's analysis is that although the two theoretical perspectives are clearly distinct, in terms of how they view the relationship between language and mind, and the claims they might make about this, she concludes

that "the inferences one anticipates drawing from verbal protocols are not dissimilar across these two theoretical perspectives" (p. 17). From this she reaches a higher generalization about the similarity between these two theoretical perspectives on the mind: "Both aim to develop claims about the higher mental processes participants make use of in carrying out a specific task" (p. 17).

Underspecification and grain-size

Two issues arise in defining constructs: underspecification and grain size. The language phenomena that we observe in applied linguistics research are rich and complex. In addition, what we observe is often but a fleeting moment in the span of someone's life of language use, or a single instance that is part of some larger phenomenon we want to describe or explain. Even though we may be able to capture this moment with a high degree of technical accuracy in video and audio media, our descriptions of the phenomenon will necessarily be limited in two ways. First, the observer may not perceive all of the elements of the speech event, no matter how many times he replays the video. Even if he were to perceive all the physical details, he may still miss the significance of some of these details as they are perceived by and affect the participants in the speech event. Because of this, any constructs that the researcher may induce from his observations may be underspecified, or incomplete, to some degree. That is, the researcher may unwittingly omit or overlook a particular feature that is of relevance to the description or explanation. Similarly, when a rater assigns a score to a test taker's performance on a language assessment, she may inadvertently miss some critical aspect of that performance that might warrant a different score. Another rater might pick up on this aspect, while missing or ignoring other features of the performance.

Second, in much research, the researcher chooses to be selective, attending only to those features that he believes or hypothesizes to be distinctive in a way that is relevant to his research questions. Thus, in order to focus on the aspects of the phenomenon that are of interest, the researcher typically makes certain simplifying assumptions that determine, to a large degree, what he observes and when, where and how he makes his observations. When a conversation analyst or ethnographer chooses a particular speech event or community to observe, or when a functional grammarian chooses a particular linguistic structure to examine, and selects a particular linguistic corpus or elicitation procedure, she is selecting or simplifying the range of phenomena to be included in the research. Similarly, when a researcher extracts one or more scores from an individual's performance on a language assessment, he is focusing only those specific aspects that performance that will enable him to make the kinds

of interpretations about ability that he is interested in, and hence simplifying the range of language performance he needs to observe. Another example of selectivity, or simplification, is when the researcher places the limitations of experimental design on the phenomena and artificially manipulates the treatment, or experience, which research participants undergo. I thus would argue that virtually all research in applied linguistics involves selectivity or simplification, so that our descriptions of phenomena are necessarily incomplete and underspecified to some degree.

A second issue is that of grain-size, or the level of detail we capture in our descriptions of phenomena. How much detail does our description of the speech event need to include? If, for example, a researcher is interested in describing the acoustic features of speech, she will need a very high quality recording and precise instruments to display the acoustic characteristics of the sound waves that are generated in the speech event. Here, she may need to measure time in a way that is accurate to nanoseconds. Another researcher, on the other hand, might be interested in doing a conversational analysis of speech. This researcher can most likely forgo an acoustic analysis of the sound waves, and may find that timings in seconds are sufficiently accurate, but will nevertheless want a very accurate transcription of the speech, including pauses and non-verbal utterances. A third researcher might be interested in understanding the strategies that a test taker reports using while taking a speaking test. In this case, she might not transcribe the speech at all, but might make notes about what processes or strategies the test taker reports using. To give another example, if a researcher wanted to know how well a student can use grammar in writing, it might be sufficient to obtain a single rating of grammar based on a written composition. If, on the other hand, he wanted to diagnose areas of strength and weakness in grammatical knowledge, he would need to obtain more specific information about different aspects of grammatical knowledge, such as knowledge of propositions, tense and aspect markers, or agreement in grammatical gender. Thus, the grain-size – the amount and level of detail – that is included in the researcher's description will depend, to a large extent, on how broadly or narrowly he defines the construct.

The nature and role of context

"Context" is a comfortable, albeit slippery term in applied linguistics;[10] we all use the term, and know what it means, yet we don't all agree on precisely what, where or even when it is. If the researcher adopts a monist ontology, then the notion of context becomes largely irrelevant, since the observer, the entity, or

construct to be described, induced or explained, and the context are all part of the same phenomenon, or reality, and it makes little sense to try to delineate where one ends and the other begins.[11] This monist view can be illustrated as in Figure 2 below.

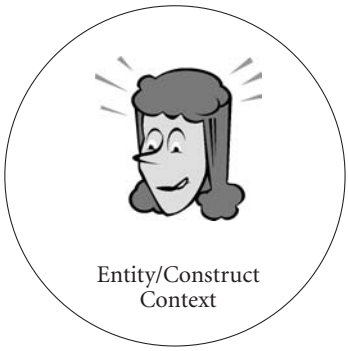

Figure 2. Monist view: Observer, the entity/construct and the context all part of phenomenon

It is this ontological view that is implicit, I believe, in ethnographic research that aims at providing a rich description, or ethnography, of the phenomenon, and takes an "emic" view of this. In such research the entity/construct and the context are essentially indistinguishable, and the researcher's role is that of "insider", or participant, who is part of the phenomenon to be described. This view seems to be implicit in much research in sociolinguistics, and is the source of the observer's paradox, the methodological conundrum discussed by Labov (1990).

Much research in applied linguistics, however, is conducted within a dualist ontology, so that the observer is essentially extracted from the context and the entity that is of interest. In this view, the researcher makes observations of the phenomenon as it happens in an objective reality in order to come to some understanding of the entity of interest. Distinguishing the entity/construct from the context, however, as well as delineating the relationship between these, becomes potentially problematic. One often-cited approach to describing context is Hymes' (1974) well-known "SPEAKING" acronym for remembering the features of a speech event (Setting, Participants, Ends, Act sequence, Key, Instrumentalities, Norms, Genres). In the area of language assessment, Bachman and Palmer (1996) have proposed a set of characteristics that can be used to describe language use tasks and language assessment tasks. Neither of these sets of

characteristics, however, recognizes the dynamic, changing nature of context, as described by Erickson and Schulz (1981). From these perspectives, context is essentially the external setting, or situation, in which the phenomenon takes place, and in which the observer observes it. One of the most inclusive definitions of context I have seen is that of research in language education settings: "the larger sociopolitical or historical context (where relevant), the participants and their interests, the tasks or instructional practices used and the participants' understandings or views of these, in some cases, and how the research itself, whether inside the classroom or in a research office of some kind, creates a special sociolinguistic context, system or ecology that is temporally as well as socially and discursively situated" (Duff this volume, p. 12, citing van Lier 1988, 1997). In a dualist ontology, context may thus be viewed as relatively static, or as dynamic and changing, but it is nevertheless clearly distinct from the observer. Within this dualist perspective, in whatever way we may choose to define context for a particular study, the critical issue that needs to be addressed is that of the relationship between that context and the entity to be described, induced or explained.

What is the relationship between the context and the entity/construct?
One dualist view of the relationship between the entity/construct and the context is that these are essentially inseparable. In this view, the researcher treats the language use activity or speech event as a unitary construct, describing both what the language users say and do and the features of the context or situation in which the speech event occurs. This view is illustrated in Figure 3 below.

This ontological view that is implicit, I believe, in much ethnographic research that investigates the ways in which language functions in the socialization of children, and in the creation of a sense of community, through the co-construction of discourse. In such research, the entity and the context are

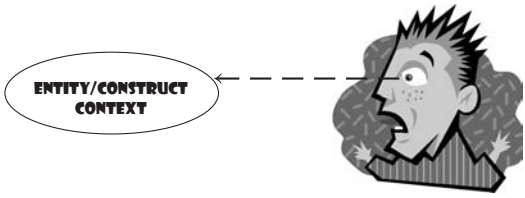

Figure 3. Dualist view: Observer independent of the phenomenon; entity/construct and context not distinguished

essentially indistinguishable, and the researcher's role is that of observer of the phenomenon (e.g., Ochs 1988; Ochs & Capps 2001).

Another dualist view would be to treat the entity/construct and the context as separate and essentially unrelated, in which case the research focuses solely on describing the entity/construct itself, independently of the context. This is the perspective that characterizes "trait" approaches to defining the construct. This view is illustrated in Figure 4 below.

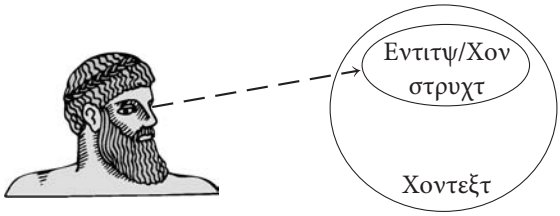

Figure 4. Dualist view: Entity/construct exclusive of context

An example of this would be giving a language assessment or eliciting grammaticality judgments in which the tasks are entirely decontextualized, and then scoring the research participants' performances on these tasks and interpreting these scores independently of either the context in which they have responded to the task, or the language use context or linguistic structure to which the researcher wants to generalize.

A much more typical dualist view in applied linguistics research, I believe, is an interactionalist one. This view entails the presupposition that the entity/construct that is of interest exists or happens within a context, and that these interact in ways that mutually affect each other. In this view, it is the interaction itself that is the entity of interest. This dualist interactionalist view is illustrated in Figure 5 below.

An example of this interactionist view, I believe, is Swain's (this volume) analysis of verbal protocols, which, she concludes, need to be considered as part of the phenomenon – the interactions among participants in the speech event and the cognitive changes that these bring about. This perspective also underlies the interaction hypothesis in SLA research (discussed by Markee this volume). Another example of an interactionalist perspective on context is that proposed by Douglas and Selinker (1985), who have argued that in addition to context as an external milieu in which speech events happen, and with which individual language users interact, language users utilize what Douglas

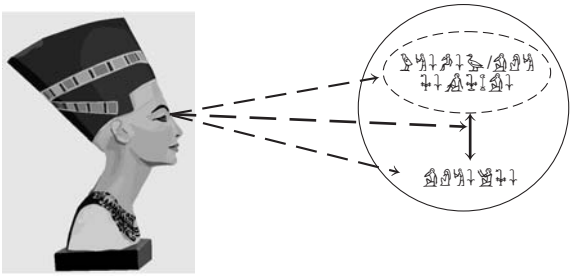

Figure 5. Dualist interactionalist view: Entity/construct and context separate but interact with each other

and Selinker call discourse domains, which are essentially personalized internal contexts that are created through language use.

The way a researcher views the relationship between the entity of interest and the context will be influenced by the purpose of her research, her ontological view and her approach to construct definition, as discussed above. These are illustrated in Figure 6.

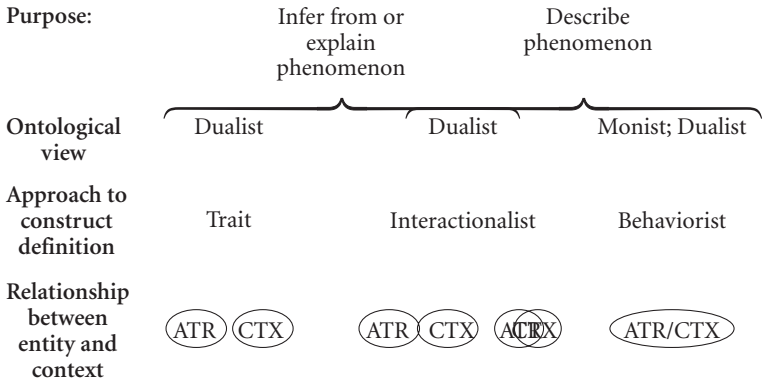

Figure 6. Differing purposes and approaches to construct definition (ATR – attribute; CTX – context)

A trait approach to construct definition, illustrated in the leftmost part of Figure 6, implies a dualist ontology, in which the researcher observes consistencies in observations as indicative of an attribute (e.g., language ability) that is a characteristic of individual language users, and that is essentially independent of the context in which language use takes place. This is typically associated

with the purpose of inferring from or explaining or predicting phenomena. As pointed out by Deville and Chalhoub- Deville (this volume) and McNamara (this volume), this clear distinction between individual ability and context has been the dominant approach, historically, for defining constructs in the area of language assessment, where the purpose is typically to explain test performance in terms of underlying abilities, or to predict some future performance.

In a behaviorist approach to construct definition, illustrated by the rightmost part of Figure 6, the attribute and the context are indistinguishable, both from each other, and from the observer. This approach could imply either a monist or a dualist ontology. The "emic" perspective in ethnographic research, I believe, provides an example of a monist ontology underlying a behaviorist approach to construct definition. In this case, the phenomenon that is observed includes all of the interactions among the participants, including those with the researcher as participant-observer, and these interactions are situated in and part of a context. When the researcher steps back and takes the "etic", outsider's stance, on the other hand, this implies a dualist ontology. In this case, a dualist ontology, with the observer observing an objective reality in which the attributes of language users and context are seen as a unitary phenomenon, could underlie a behaviorist approach to construct definition. Conversational analysis, I believe, provides an example of a dualist ontology underlying a behaviorist approach.

The last approach to construct definition, the interactionalist, illustrated in the middle part of Figure 6, is, I believe, the most recent to emerge in applied linguistics research, and is thus the one that is still the least well-problemmatized. This approach implies a dualist ontology, but unlike the other two approaches, which view attributes of language users and contexts as either completely separate and independent or as essentially indistinguishable, the interactionalist approach treats attributes and contexts as distinguishable but interacting entities. The extent to which attributes and contexts are seen to interact may differ from one researcher to another, or from one study to another, with the attribute and context interacting very little in some cases and a great deal in others. As has been pointed out by Chalhoub-Deville (2003), although an interactionalist approach to defining language ability is theoretically compelling, in terms of its recognition of language use as co-constructed through social interaction, it is not without its challenges, in terms of real-world application. In language assessment, she discusses as a challenge "the notion that language ability is local and the conundrum of reconciling that with the need for assessments to yield scores that generalize across contextual boundaries" (Chalhoub-Deville 2003: 373). I would argue that this is also a challenge for

any research in applied linguistics in which the purpose is to generalize beyond the specific observations that are part of the research.

What is the range or scope of the context?

Whatever perspective on context the researcher may adopt, for any given study he will generally need to focus, or delimit, the range or scope of the context he intends to observe. Duff (this volume), for example, points out that "in case studies and classroom ethnographies, it is necessary for researchers to acknowledge the delimited context- (or culture-) bounded nature of their observations" (p. 11). In her examples, the scope of the context ranges from the individual language learner to the entire classroom, to extracurricular activities. In conversational analysis, the scope of the context may be a single conversation, or a single adjacency pair. Larsen-Freeman (this volume) gives an example of an investigation of Wh-clefts in which the contexts included face-to-face conversations, telephone conversations, talk radio broadcasts, group therapy sessions, and a 1.8 million word computerized corpus of spoken academic discourse, containing an amalgam of contexts. For Swain (this volume) there are two contexts: (1) the collaborative dialogue that takes place between two language learners and (2) their subsequent interaction with the researcher as they verbalize in stimulated recalls. In other areas of applied linguistics research, such as the evaluation of language programs, language planning, or ethnography, the context may be an entire educational program or a country. Contexts in applied linguistics research thus cover a vast range, in terms of their scope.

The observation and report

What counts as an "observation"?

As has been suggested above, the phenomena we observe in empirical research in applied linguistics are varied, as are the ways in which we observe them. Thus, we might well ask what distinguishes the observations we make in research from the casual observations of the lay person. As I have argued elsewhere (Bachman 2004b), I believe that our observations in research are distinguished by two characteristics: (1) they are systematic and (2) they are substantively grounded. Observations in empirical research are systematic in that they are planned and carried out in a way that is clearly described and thus accessible to other researchers and consumers of research. In other words, the observations are conducted according to explicit procedures that are accepted and current practice in a particular field, and that are open to scrutiny

by other researchers. Observations that we make as part of research are also substantively grounded, in that they are related to or based on accumulated knowledge from prior research or a widely accepted theory about the nature of the phenomenon we want to describe or explain. The observations reported in the chapters in this volume vary from elicited responses, such as grammaticality/acceptability judgments (Larsen-Freeman) or test scores (Chapelle, Deville & Chalhoub-Deville, McNamara); and stimulated recall verbal protocols (Swain), to "naturally occurring" events, such as conversations (Markee), classroom interactions (Duff) and language corpora (Larsen-Freeman).

What is the unit of analysis?
Related to the issue discussed above, about grain-size, is what the researcher identifies as the unit of analysis. If a researcher wanted to investigate the relationship between English language learners' (ELLs) scores on a classroom assessment and a standardized test of English proficiency, the scores received by the individual students might be the unit of analysis. However, recognizing that teachers and classroom interactions may affect ELLs' English proficiency, it might be of interest to aggregate their scores within classrooms to investigate these effects, in which the teacher/classroom becomes the unit of analysis (e.g., Llosa 2005). In conversational analysis, the unit of analysis might be a single utterance, an adjacency pair, or an entire conversation, while in ethnographic research the unit of analysis might be a group of language users involved in a speech event, the interactions between students and a teacher in a classroom, or the interactions that take place among members of a family or an entire community.

How is the observation reported?
Researchers may report their observations in a number of ways. If we give a group of students a test, for example, we will report the results as scores, with perhaps a separate score for each part of the test. In much research in applied linguistics, however, the reports of research consist of words, or verbal descriptions of what the researcher has observed. Researchers also use pictures, charts, audio and video recordings to report their observations.

The observer's interpretation of the observation

As illustrated in Figure 1 above, the researcher's interpretation of the phenomenon is based on some performance that is observed and the report of that observation. The way the researcher interprets the observation will depend on

her ontological stance toward the research. The credibility, trustworthiness, or validity of the interpretation will depend on the cogency of the researcher's argument linking the observation and the interpretation, and the persuasiveness of the evidence the researcher is able to marshal in support of that argument.

What is the researcher's ontological stance towards the interpretation?
In research whose purpose is to "draw inferences from the data collected" (Swain this volume), or to "generate new insights and knowledge" (Duff this volume) on the basis of observations, the way the researcher views his interpretation and the knowledge he claims to have acquired through the research, will depend on his ontological stance. The researcher's ontological "stance" is the status the researcher accords to his interpretation of his observations in a particular study and what this represents, in terms of knowledge acquired.[12] In other words, the researcher's ontological stance is the way he views the knowledge that is acquired on the basis of his research. Three different ontological stances are relevant to research in applied linguistics: operationalist, realist and constructivist.

Operationalist stance. In an operationalist stance, the meaning of the construct is essentially synonymous with the operations or procedures that are used to observe the phenomenon. That is, the construct is defined in terms of the method of observation, and "meaning is constituted in empirical operations" (Bickhard 2001: 2). In language assessment, an operationalist stance would define the construct, "language proficiency", for example, as whatever a particular test of this measures. One of the logical consequences of an operationalist stance is that each observation defines, in effect, a different construct. Despite the fact that two forms of a reading comprehension test, for example, might present test takers with very similar reading passages and ask the same kinds of comprehension questions, the operationalist would claim that the abilities measured by these two tests are essentially different constructs. An operationalist stance is thus clearly problematic if the researcher's purpose is to generalize beyond the particular observation, to either another group or context, or to attaching meaning to the observation in terms of an abstract construct. However, if the purpose of the research is to provide a rich description of the phenomenon, then this could imply, essentially, an operationalist stance.

Realist stance. A realist stance holds that the phenomenon we observe exists independently of the researcher herself and what she may believe or theorize about the phenomenon. According to this stance, the product of empirical re-

search is knowledge of largely theory-independent phenomena. Furthermore, such knowledge is possible and actual, even in those cases in which the relevant constructs are not directly observable (Boyd 2002). Realism about the everyday world entails two claims: existence and independence. The realist claims that objects such as tables, rocks and the moon, all exist, as do the facts that a particular table is square, a rock is granite and the moon is spherical. The second claim is that these objects in the real world and their properties (e.g., squareness, graniteness, sphericity) are independent of what the researcher may happen to think or say about them (Miller 2004).

Constructivist stance. The constructivist stance regards the inference or interpretation as nothing more than a construction of the researcher's mind, which may or may not be independent of the observation (Boorsboom 2003: 207).[13] Thus, the constructivist rejects the realist's claim about the nature of the construct. For the constructivist, the knowledge that we acquire through research is neither "real" in the realist sense, nor fixed, but is constructed by the researcher through her own observation of the phenomenon. Unlike the realist, who may be searching for a single, "objective", "correct" interpretation of the observed phenomena, the constructivist admits to, and may even seek, multiple perspectives on or constructions of the phenomenon. It is important not to confuse the constructivist stance with the way in which the researcher arrives at or constructs an interpretation. In virtually all applied linguistics research, the interpretations are constructed by the researcher, even though he may believe that his interpretations are about phenomena that exist in the real world. Similarly, a constructivist stance should not be identified with the way the researcher views the phenomenon to be observed. Thus, many researchers in applied linguistics view the phenomenon – speech event, conversation – as constructed through interaction, but still view these as entities that exist in the real world.

In my view, all of the chapters in this collection adopt an essentially realist stance. Chapelle, McNamara and Deville and Chalhoub Deville, while they may not agree on exactly where the construct resides (in language users, in contexts, or in the interactions between language users and contexts), all discuss the constructs that they infer on the basis of language assessments in realist terms. Chapelle discusses constructs such as "vocabulary size" and "organization of L2 lexicon", and makes repeated references to a "mental lexicon", which I interpret as a claim that these constructs have some reality in the minds of test takers. McNamara also appears to adopt a realist stance, discussing constructs, initially in terms of what individual learners know and can do in a criterion domain,

and as cognitive and social characteristics of test takers (p. 6). He then discusses constructs as co-constructions within oral discourse, and as "an expression of the outcomes based and functionalist demands of current government policy on adult education and training" (p. 38). All of these – performance in criterion domains, characteristics of test takers, oral discourse, and government policy – presumably exist in a real world. Similarly, Deville and Chalhoub-Deville initially discuss language ability as "a stable construct residing within individuals" (p. 12), and then discuss it as "ability-in-language user-in context" (p. 18), which resides in the interactions between language users and the context. Again, I infer that these constructs within individuals, and these interactions in contextualized language use exist in a real world. While Markee and Swain both clearly view the phenomena they observe (conversations and stimulated recalls, respectively) as constructed, the constructs (conversational repairs and thought, or cognitive processes, respectively) that they infer from their observations imply a realist ontological stance. For Markee, the conversational repair is a real feature of human conversation; for Swain, thought and cognition exist somewhere in language users. Larsen-Freemen infers constructs (e.g., the form, meaning and use of the WH-cleft in English spoken discourse) from a variety of phenomena (prior research, intuition, observed language use, linguistic corpora), and seems to adopt the realist stance that these constructs, or patterns exist in the language use of English speakers. Duff constructs interpretations on the basis of her observations or perceptions, which might not be the same as other researchers' perceptions and interpretations. Nevertheless, she views the constructs she infers from her observations as existing in the real world, independent and uninfluenced by her own perceptions of them.

What is the relationship between observations and interpretations?

The inferential links between the observed performance, the observation results, the interpretation, and the use of the research results were discussed above, and are illustrated in Figure 1. But while the researcher may claim that these links are justified, if she wants to convince the various audiences to whom the research is directed, the inferences that constitute these links need to be supported by a coherent logical argument and evidence. While the specific details of that argument and the kinds of evidence that are acceptable, or convincing, varies across the different research approaches in applied linguistics, I would propose that the basic *structure* of these arguments is essentially the same, and can be characterized more generally as a "research use argument".

Bachman (2005) has argued that in order to support the interpretations and uses we make of individuals' performance on language assessments, we

need to articulate an assessment use argument. Such an argument provides a logical framework for linking performance on language assessments to intended interpretations and uses, and can be used not only as a guide in the design and development of language assessments, but can also inform a program of research for collecting the most critical evidence in support of the interpretations and uses for which the assessment is intended. As described by Bachman, the structure of an assessment use argument follows Toulmin's (2003) argument structure. At the center of an assessment use argument is a link between some *data*, or observations, and a *claim*, or an interpretation of those observations. This link is supported by various *warrants*, which are statements that we use to justify the link between the observation and the interpretation. These warrants are in turn supported by *backing*, which consists of evidence that may come from prior research or experience, or that is collected by the researcher specifically to support the warrant. In addition to these parts, the argument also includes *rebuttals*, or potential alternative interpretations or counterclaims to the intended interpretation. These rebuttals may be either weakened or strengthened by *rebuttal data* (Bachman 2005: 10).[14]

I would suggest that a similar "research use argument" can provide a framework for guiding empirical research in applied linguistics. The primary function of a research use argument would *not* be to falsify a theory, in a positivistic sense, but rather to convince or persuade a particular audience – fellow researchers, journal reviewers, funding agencies, tenure committees, or consumers of the research – that the researcher's claim or interpretation is useful for some purpose. To paraphrase Chapelle (this volume), I believe that many of the problems in applied linguistics can best be addressed by research that is seen as "true enough to be useful" (p. 1). The usefulness of research in applied linguistics should be judged, in my view, not by the extent to which it captures a glimpse of the "Truth", but by the cogency of the research use argument that underlies the research and the quality of the evidence that is collected by the researcher to support the claims made in the argument. I believe that using such an argument as a rationale and organizing principle for research would enable researchers in applied linguistic to break away from our attempts to emulate "scientific" research in the physical sciences, and from the never-ending paradigm debate between the so-called quantitative and qualitative approaches to research. Indeed, I believe that a research use argument can provide a logical framework and rationale for the appropriate use of multiple approaches, or mixed methods, for collecting evidence to support the claims made by the researcher.

Combining different perspectives and approaches in research

Is the complementary use of multiple approaches in a single study desirable or even possible? What is the value added of attempting to use complementary approaches? Is it possible to gain important insights about a phenomenon through multiple approaches? As McNamara (this volume) points out, there are numerous examples of research in the area of language assessment that have productively combined qualitative and quantitative approaches to yield richer and, in my view, better supported interpretations and insights into the phenomena they have investigated.[15] As mentioned above, the metaphor of triangulation is often used to describe a viable way to obtain converging evidence from multiple approaches. However, multiple approaches can also be used to investigate possible alternative interpretations, or what Lynch (1996) calls "negative case analysis", the search for instances in the data that do not fit with an interpretation suggested by other data in the study (p. 57). Thus, with respect to using multiple approaches, we need to ask, "at what level are the approaches combined – philosophical perspective, approach to defining the phenomenon or construct that is of interest, procedures for observing and reporting, the researcher's ontological stance?" In many studies, it is not always clear that there is a genuine combining of approaches; rather, the combination appears to be opportunistic and unplanned. For example, quantitative research with a qualitative "add on" to hopefully help make some sense of the quantitative analyses, or counting frequencies of occurrences of categorizations based on naturalistic observation, and then using a test of statistical significance, might superficially combine aspects of different research approaches, without integrating these into a single coherent study. There is thus a need for a principled basis for determining, for a particular study, whether to combine approaches, and if so, at what level, how, and why.

There is a rich literature, going back to the mid 1980's, in the social sciences and education, discussing multiple approaches, or what is more commonly referred to as "mixed methods" research, and several collections and handbooks of mixed methods research are available (e.g., Brewer & Hunter 1989; Creswell 2003; Johnson & Christensen 2004, 1998; Tashakkori & Teddlie 2003). (See also www.ehr.nsf.gov/EHR/REC/pubs/NSF97-153/START.HTM). Recently, Johnson and Onwuegbuzie (2004) have argued for a pragmatic and balanced or pluralist position to inform the use of qualitative, quantitative or mixed approaches to research. Drawing on the work of the American pragmatic philosophers Peirce, James and Dewey, they arrive at a pragmatic principle that is essentially the same as the criterion of the usefulness of research that I have

discussed above: "when judging ideas we should consider their empirical and practical consequences" (Johnson & Onwuegbuzie 2004: 17.) Extending this principle to research methods, they argue that the decision on what research approach to take – quantitative, qualitative or mixed – for a given research study should be based not on doctrinaire or philosophical positions, but rather on the consideration of which approach is the most appropriate and likely to yield the most important insights into the research question.

> Consideration and discussion of pragmatism by research methodologists and empirical researchers will be productive because it offers an immediate and useful middle position philosophically and methodologically; it offers a practical and outcome-oriented method of inquiry that is based on action and leads, iteratively, to further action and the elimination of doubt; and it offers a method for selecting methodological mixes that can help researchers better answer many of their research questions.
>
> (Johnson & Onwuegbuzie 2004: 17)

Johnson and Onwuegbuzie go on to describe a typology of mixed research methods, along with a process model for making decisions about research methodologies in the design and implementation of research.

Empirical research in applied linguistics has, I believe, much to learn from the literature on mixed methods research. As the examples of mixed methods research from language assessment that McNamara (this volume) discusses, it is clear, that such research is not only feasible, but is also highly desirable, in that it expands the methodological tools that are at the researcher's disposal, thereby increasing the likelihood that the insights gained from the research will constitute genuine advances in our knowledge.

Conclusion

I have taken the opportunity presented in this response chapter to outline a broad view of generalizability in empirical research in applied linguistics. According to this view, generalizability is a concept that has different aspects – consistency, meaningfulness and consequences – that provide empirical researchers in applied linguistics a basis for interpreting their observations, and for considering the possible uses and consequences of these interpretations. I have suggested that we adopt an epistemology of argumentation that moves away from one that seems to be driven by a preoccupation with our differences (e.g., quantitative, positivistic, theory falsification vs. qualitative, naturalistic,

descriptive; "science" vs. "non-science"), and toward one that admits a wide range of empirical stances and methodologies. Many of the issues that I have touched on here, such as the mind-body problem and monism vs. dualism, have perplexed philosophers and scientists for centuries, and continue to be energetically debated widely in philosophy, the social sciences, cognitive science and neuroscience, to this day. Many researchers in our field have also discussed issues related to epistemology and the goals of research (e.g., Gregg 1993; Jordan 2004; Lantolf 1996; Long 1990, 1993; Schumann 1984; Schumann 1993). I would thus be the first to say that my discussion of these issues here is by no means either authoritative or conclusive. Indeed, much of what I have discussed may be of little interest or concern to many very productive and creative researchers in our field. However, having recently co-taught a graduate proseminar in epistemology in applied linguistics, I am only too painfully aware of how uninformed many of our students are about these issues. Thus, my primary reason for discussing such arcane topics as monism and dualism, trait, behaviorist and interactionalist perspectives, and operationalist, realist and constructivist ontological stances, is that without an awareness of these issues, and of the alternatives in empirical research they present, it is difficult for researchers to apply the kind of critical self-reflection on their work that I believe is essential for the health and vitality of the field. Similarly, an uncritical application of one particular research methodology, without an understanding of the ontological and epistemological assumptions on which it is based, as well as the possible consequences of the outcomes, can, I believe, lead to students' robot-like emulation of the research of senior scholars and mentors, with little of their creativeness and insight.

Notes

* I would like to thank Carol Chapelle, Patsy Duff, Lorena Llosa and Adrian Palmer for their helpful comments and suggestions on earlier drafts of this chapter.

1. See Bachman (2005) for a discussion of the logical structure of an assessment use argument.

2. As discussed below, the "construct" is what the researcher wants to describe, infer or explain on the basis of observing some phenomenon.

3. See Heider (1988) for a discussion of different ways of interpreting differences among ethnographies, and factors that may contribute these.

4. See Bachman (2004b) for a discussion of procedures and approaches to validation in language assessment.

5. For an excellent discussion of validity in both quantitative and qualitative research, see Lynch (1996), from which I have drawn in my discussion of validity in qualitative research.

6. See Chapelle (1998) for a discussion of the social consequences of using language tests in SLA research.

7. Discussions of ethics and consequences in language testing can be found in the following edited collections (Davies 1997b, 2004; Kunnan 2000).

8. The debate about the mind-body problem has produced a wide variety of specific monist and dualist perspectives, and a discussion of this issue and these perspectives is far beyond the scope of this chapter.

9. For an excellent set of discussions of the issues surrounding the emic-etic distinction, see Headland (1990).

10. Chalhoub-Deville and Deville (forthcoming) also characterize context as a "slippery issue".

11. In this section, I use the term "entity" and "construct" to distinguish this from the context. The entity or construct can consist of a trait, or attribute of the language users, of the phenomenon itself (attributes and context not distinguished), or of the interaction between attributes of the language users and features of the context.

12. See Borsboom, Mellenbergh and van Heerden (2003) for an excellent discussion of these stances in the context of latent variable models in psychology.

13. The constructivist ontological stance with respect to the acquisition of knowledge through empirical research draws on the same philosophical perspective as do constructivist views of learning.

14. See Koenig and Bachman (Koenig & Bachman 2004) for a discussion of how an assessment use argument could inform the investigation of the validity of interpretations based on accommodated standardized achievement tests. See Llosa (2005) and Ockey (2005) for applications of an assessment use argument to validation studies in language assessment.

15. In addition to the studies cited by McNamara, I would point to Weigle (1994, 1998), Sasaki (1996, 2000), and Sawaki (2003).

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: OUP.

Bachman, L. F. (2004a). Linking observations to interpretations and uses in TESOL research. *TESOL Quarterly, 38*(4), 723–728.

Bachman, L. F. (2004b). *Statistical analyses for language assessment*. Cambridge: CUP.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: OUP.

Bickhard, M. H. (2001). The tragedy of operationalism. *Theory and Psychology, 11*(1), 35–44.

Boorsboom, D. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203–219.

Boyd, R. (2002). Scientific realism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (*Summer 2002 Edition*): http://plato.stanford.edu/archives/win2004/entries/scientific-realism/.

Brewer, J. & Hunter, A. (1989). *Multimethod research: A synthesis of approaches*. Newbury Park, CA: Sage.

Campbell, D. T. & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.

Chalhoub-Deville, M. & Deville, C. (forthcoming). Old, borrowed, and new thoughts in second language testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). New York: American Council on Education and Macmillan Publishing Company.

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford: OUP.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). New York, NY: CUP.

Chapelle, C. A. (this volume). L2 vocabulary acquisition theory: The role of inference, dependability and generalizability in assessment. In M. Chalhoub-Deville, C. A. Chapelle, & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives*. Amsterdam: John Benjamins.

Churchland, P. M. (1996). *The engine of reason, the seat of the soul*. Cambridge, MA: The MIT Press.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed approaches*. Thousand Oaks, CA: Sage.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason and the human brain*. New York, NY: Grosset/Putnam.

Davies, A. (1997a). Demands of being professional in language testing. *Language Testing, 14*, 328–339.

Davies, A. (1997b). Special issue: Ethics in language testing. *Language Testing, 14*(3).

Davies, A. (2004). Special issue: The ethics of language assessment. *Language Assessment Quarterly, 1*(2/3).

Dennett, D. C. (1991). *Consciousness explained*. New York, NY: Little, Brown & Co.

Deville, C. & Chalhoub-Deville, M. (this volume). Old and new thoughts on test score variability: Implications for reliability and validity. In M. Chalhoub-Deville, C. A. Chapelle & P. A. Duff (Eds.), *Inferences and generalizability in applied linguistics: Multiple perspectives*. Amsterdam: John Benjamins.

Douglas, D. & Selinker, L. (1985). Principles for language tests within the 'discourse domains' theory of interlanguage: Research, test construction and interpretation. *Language Testing, 2*, 205–226.

Duff, P. A. (this volume). Beyond generalizability: Contextualization, complexity, and credibility in applied linguistics research. In M. Chalhoub-Deville, C. A. Chapelle, & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives*. Amsterdam: John Benjamins.

Erickson, F. & Schulz, J. (1981). When is context? Some issues in the analysis of social competence. In J. Green & C. Wallat (Eds.), *Ethnography and language in educational settings*. Norwood, NJ: Ablex.

Gregg, K. R. (1993). Taking explanations seriously; or, let a couple of flowers bloom. *Applied Linguistics, 14*, 276–294.

Guba, E. G. & Lincoln, Y. S. (1982). Epistemological and methodological bases of naturalistic inquiry. *Education Communication and Technology, 30*(4), 233–252.

Hamp-Lyons, L. (1997a). Ethics in language testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (Vol. 7: Language testing and assessment, pp. 323–333). Dordrecht: Kluwer Academic.

Hamp-Lyons, L. (1997b). Washback, impact and validity: Ethical concerns. *Language Testing, 14*, 295–303.

Hart, W. D. (1988). *The engines of the soul*. Cambridge: CUP.

Headland, T. N., Pike, K. L., & Harris, M. (Eds.). (1990). *Emics and etics: The insider/outsider debate*. Dallas, TX: Summer Institute of Linguistics.

Heider, K. G. (1988). The Rashomon effect: When ethnographers disagree. *American Anthropologist, 90*(1), 73–81.

Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia, PA: University of Pennsylvania Press.

Johnson, R. B. & Christensen, L. B. (2004). *Educational research: Quantitative, qualitative and mixed approaches*. Boston, MA: Allyn and Bacon.

Johnson, R. B. & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Jordan, G. (2004). *Theory construction in second language acquisition research*. Amsterdam: John Benjamins.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(3), 135–170.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational measurement: issues and practice, 18*(2), 5–17.

Kirk, J. & Miller, M. L. (1999). *Reliability and validity in qualitative research*. Newbury Park, CA: Sage Publications.

Koenig, J. A. & Bachman, L. F. (Eds.). (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessment*. Washington, DC: National Research Council, National Academies Press.

Kunnan, A. J. (Ed.). (2000). *Fairness and validation in language assessment: Selected papers from the 19th language testing research colloquium, Orlando*. Cambridge: CUP.

Labov, W. (1990). *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania Press.

Lantolf, J. P. (1996). Second language theory building: Letting all the flowers bloom. *Language Learning, 46*, 713–749.

Larsen-Freeman, D. (this volume). Functional grammar: On the value and limitations of dependability, inference, and generalizability. In M. Chalhoub-Deville, C. A. Chapelle, & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives*. Amsterdam: John Benjamins.

Llosa, L. (2005). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency. Unpublished PhD dissertation, University of California, Los Angeles.

Long, M. H. (1990). The least a second language acquisition theory needs to explain. *TESOL Quarterly, 24*, 649–666.

Long, M. H. (1993). Assessment strategies for SLA theories. *Applied Linguistics, 14*, 225–249.

Lynch, B. K. (1996). *Language program evaluation: Theory and practice*. Cambridge: CUP.

Markee, N. (this volume). A conversation analytic perspective on the role of quantification in second language acquisition research. In M. Chalhoub-Deville, C. A. Chapelle, & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives*. Amsterdam: John Benjamins.

Mathison, S. (1988). Why triangulate? *Educational Researcher, 17*(2), 13–17.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review, 62*(3), 279–299.

McNamara, T. (this volume). Validity and values: Inferences and generalizability in language testing. In M. Chalhoub-Deville, C. A. Chapelle, & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives*. Amsterdam: John Benjamins.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89*, 575–588.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3 ed., pp. 13–103). New York, NY: American Council on Education and Macmillan Publishing Company.

Miller, A. (2004). Realism. In E. N. Zalte (Ed.), *The Stanford encyclopedia of philosophy* (*Winter 2004 Edition*): http://plato.stanford.edu/archives/win2004/entries/realism/.

Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability and Risk, 2*, 237–258.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 47–72). New York, NY: Academic Press.

Ochs, E. (1988). *Culture and language development: Language acquisition and language socialization in a Samoan village*. Cambridge: CUP.

Ochs, E. & Capps, L. (2001). *Living narrative*. Cambridge, MA: Harvard University Press.

Ockey, G. J. (2005). DIF techniques to investigate the validity of math word problems for English language learners. PhD Qualifying Paper, University of California, Los Angeles.

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: Quantitative and qualitative analyses*. New York, NY: Peter Lang.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing, 17*(1), 85–114.

Sawaki, Y. (2003). A comparison of summary and free-recall as reading comprehension tasks in a web-based placement test of Japanese as a foreigh language. Unpublished PhD dissertation, University of California, Los Angeles.

Schumann, J. H. (1984). Art and science in second language acquisition research. In A. Z. Guiora (Ed.), *An epistemology for the language sciences* (*language learning, special issue*) (pp. 49–76). Ann Arbor.

Schumann, J. H. (1993). Some problems with falsification: An illustration from SLA research. *Applied Linguistics, 14*, 295–306.

Swain, M. (1993). Second language testing and second language acquisition: Is there a conflict with traditional psychometrics? *Language Testing, 10*(2), 193–207.

Swain, M. (this volume). Verbal protocols: What does it mean for research to use speaking as a data collection tool? In M. Chalhoub-Deville, C. A. Chapelle, & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives*. Amsterdam: John Benjamins.

Tashakkori, A. & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Tashakkori, A. & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Toulmin, S. E. (2003). *The uses of argument* (Updated ed.). Cambridge: CUP.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197–223.

Weigle, S. C. (1998). Using facets to model rater training effects. *Language Testing, 15*(2), 263–267.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: CUP.

Generalizability

What are we generalizing anyway?

Susan Gass
Michigan State University

This paper pulls together strands of generalizability from the papers in the entire volume by focusing on the *why* and *what* of generalizability. The paper reflects on issues of complementary data sources and scope of inquiry, recognizing that many authors approach the *why* question from different perspectives (e.g., testing, pedagogy, acquisition). The paper is primarily concerned with issues of acquisition and raises questions relating to what we are trying to generalize and how we might know if we have been successful.

The papers in this volume approach the issue of generalization from widely different perspectives. In some respects they deal with a discussion of the virtues of one type of data over another or one approach over another. In a positive sense, this diversity helps us understand the great complexity involved in a discussion of generalization and inferences and the richness of traditions available to Applied Linguists. My own view is that there is a place for both, but we need to examine the questions that are being asked and the type of data that might or might not be obtainable to address those questions. It is a mistake for the field of applied linguistics in general or second language acquisition, in particular, to think that one approach yields "better" answers. McNamara, from a testing perspective, states this succinctly: ". . .we should not be constrained from embracing the full range of research methods open to us, be they neo-positivist or otherwise, if they deepen (as I believe they do) our ability to understand the bases for inferences in language tests" (p. 20). That is not to say that "anything goes," for it may be the case that one type of data does in fact yield a more complete picture (see Swain's chapter), but a more conservative approach is to look at all sides of the methodological issue and determine in the first instance how approaches might complement one another.

Larsen-Freeman asks the most basic question of all: why should we generalize and, given the context of applied linguistic research, what level of generalizability should we aim for? It is clear that our data need to be dependable, and it is also clear that the explanations for our data need to be generalizable. But, how do we obtain reliable and dependable data? Larsen-Freeman discusses complementary data sources. What is not always obvious, however, is which complementary data sources should be used, how they are to be gathered, and how the data should be interpreted.

As an illustration of complementary data sources, I turn to a subset of verbal protocols, known as stimulated recalls, also discussed by Swain. Verbal protocols, as Swain notes, are a type of data obtained "by asking individuals to vocalize what is going through their minds as they are solving a problem or performing a task" (Gass & Mackey 2000: 13). I choose to discuss this methodology because stimulated recall data represent a means to generalize with greater confidence beyond raw quantifiable data. To exemplify this, I turn to a recent study (Polio, Gass, & Chapin 2005) that used stimulated recall data not as an end, but as a means of supporting other data, in their specific case, interaction data. The specific research area in that study was an investigation of feedback by two groups: 10 preservice teachers and 2 experienced teachers.[1] Polio, Gass and Chapin (2005) asked whether there was a difference in the amount of implicit feedback in a dyadic task that was dependent on the amount of experience native speakers (in their case, teachers) have interacting with learners. The first round of quantified results left the researchers with the difficulty of interpreting the findings, which were not consistent with results of an earlier study by Mackey, Polio, and McDonough (2004). Generalizability and inferrability, as many papers in this volume discuss are, of course, dependent on making sure the results do in fact reflect some semblance of reality, either the reality of learners, as most work in second language acquisition, or the reality of teachers, as was the case in the Polio, Gass and Chapin study. In that study, the stimulated recall comments allowed the researchers to ask the more general question of how these two groups of teachers perceive interactions with nonnative speakers. In other words, through the recall comments, they asked if one could verify the conclusions drawn from quantitative data and thus be confident in the ability to generalize the differences found in preservice versus experienced teachers. One such difference was noted in the way the experienced teachers used language to get students to produce language. The experienced teachers did this through a variety of means. For example, they initiated the exchange by putting the burden on the student to begin the interaction, as can be seen in the examples below (ET=experienced teacher).

The primary way of doing this was by asking open-ended questions, as in (1) and (3), or by starting with an open-ended imperative as in (2).

(1) ET1: Ok we're looking at a picture. What does your picture look like?

(2) ET2: Yea. Um, well, we can kind of work together to describe the picture. So why don't you start first?
S: Oh.
ET2: Just tell me what's in your picture.

(3) ET3: So, you have a picture and I have a picture, right?
S: Yeah.
ET3: We can see our pictures, so what we have to do is to to find what's different in my picture from your picture.
S: Yeah.
ET3: There's ten differences, right?
S: Yeah.
ET3: So why don't you tell me what's in your picture? Just generally.

In these data we can only state that the teachers gave the learners an opportunity to speak, but we cannot address issues of motivation, nor can we infer what the intent of the question or imperative was. We might make a leap and infer that teachers understood the importance of output and hence wanted to provide learners with opportunities to produce language, but it is only an inference awaiting confirmation.

Following each video-taped dyadic interaction, the researchers showed each teacher (both preservice and experienced) a videotape of his/her interaction, asking for comments on particular parts of the interaction. Below are stimulated recall comments from these teachers that support the inferred interpretation of teacher behaviour.

(6) Comment, ET1: Ok, I was already thinking here, y'know, how much should I lead? Ok, should I- do I want to lead her a lot or how much will she talk? Y'know, that kinda thing.

(7) Comment, ET2: I was just trying to get her to start first so that she would take the lead in communicating.

(8) Comment, ET3: I guess I was thinking, um, I'm glad she's talkative.

The preservice teachers are quite different; their openings (9)–(12) involve specific yes/no questions and their comments (13)–(17) focus on task-related or procedural issues rather than with the need to have students produce language. This was strongly reflected in the stimulated recall comments, where they over-

whelmingly commented on the nature of the task at hand or their strategies (or lack thereof) for completing the task. A few examples from the preservice teachers are given below (PS=preservice).

(9) PS1: Um, ok. So let's start by comparing the pictures. So, is your-is your window closed?

(10) PS2: Um, is there a table in-?

(11) PS4: Well, mine looks like a picture of a dining room.
S: Um-hm.
PS4: Like with a window and a china cabinet and a picture and a stove and a rug under the table. Is that what yours kind of looks like?

(12) PS7: All right. Is your picture a picture of a kitchen?

Again, as with the experienced teachers, all we can do is describe what they are doing. That is, these preservice teachers are asking specific questions that can in fact be answered with little language.

(13) Comment, PS1: Um, yeah, at the beginning of experiment or whatever you call it . . . activity, like I wasn't sure exactly what to do, like, and, and so, like I just, just was kinda trying to figure this stuff out without really thinking about the ESL student.

(14) Comment, PS2: I remember thinking that it was a difference and that that was the first thing that we had- I feel- I felt like he finally had found a difference in the picture. Um, and then I wasn't sure if, I wasn't quite sure at first if I had to be the one to keep asking questions or if we would start asking questions, but I think, eventually, towards the end, when we got down to the last few, we were both trying to really work on it. But I think, at first, I felt like I was doing more of the asking and he was just answering.

(15) Comment, PS4: I remember thinking, like, I should probably start because I'm like, I'm the native-speaker, sort of, in this and um, she wasn't really saying anything, and I thought, since I have the paper and I tend towards leadership anyway, I'm like, ok, well I'll just start the conversation and see what, see what happens, so. . .

(16) Comment, PS7: . . . I-I remember at the beginning, it was hard to figure out where to start. That's about it.

The stimulated recall comments give us confidence to assume that language production is not the issue.

A second area where the stimulated recall comments demonstrated how the experienced teachers attempted to get the NNSs to produce language can

be seen in their use of words such as *get her/get him, have him/her, ask him/her*, as can be seen in (17). But, it is only by considering the recall comments that we can be confident in our interpretation that learner output was important to the way the experienced teachers approached the dyadic task.

(17) Emphasis on output

ET1: . . .I want to just see what she can say. . .
ET1: I should have had her asking me questions. . ..
ET2: I was just trying to get her to start first. . ..
ET6: so I was gonna to try to get him to shape-to describe
ET6: I was going to try to see if he knew the word 'cabinet'
ET7: I was just trying to get her to describe it . . .
ET7: . . .and then trying to help her get it out.
ET7: Just trying to get her to elaborate. . .

Thus, through the recall comments we can see how this group of teachers focused on the need to have learners produce language as a way of learning. This focus was not evident in the recall comments of the preservice teachers. Confidence in interpretations and hence generalizations were not possible through an inspection of the quantitative data alone. Chapelle argues, in relation to vocabulary assessment, that the crucial link is the concept of inferences, a point I return to below. "What is needed is not a single method of measurement but defensible inferences to appropriate constructs and a single framework" (p. 12). In other words, the way we make inferences, or interpret the data, is an essential ingredient of good research. In the case of stimulated recalls, the step of inferencing is reduced, if not in some cases eliminated, because we take an individual's statement as the valid interpretation of data (see Gass & Mackey 2000 for a discussion of the strengths and limitations of this methodology).

Larsen-Freeman answers her own question about the *if* of generalization and argues that generalization within limits is a necessary part of applied linguistics research ". . .I have staked out the position that generalizability is desirable – that what we would like is to have claims that apply beyond particular instances" (p. 21). It is my view that we must have the capacity to generalize, for, if not, research remains at the "that's interesting" stage rather than moving the field along in any theoretically serious way. Larsen-Freeman goes on to point out that there are limits of generalizability "at least for applied purposes." She specifically states "While some linguists pursue the broadest possible generalizations, which are therefore necessarily abstract. . .it is difficult to see how such abstract principles are useful for all the purposes to which applied linguists wish to put them" (p. 21).

This brings us to a central question of what the scope of inquiry is for which we are making generalizations. Larsen-Freeman's emphasis seems to be on generalizations about language for pedagogical purposes. Others in this volume are more focused on issues of assessment (e.g., Chapelle, McNamara, Deville, & Chalhoub-Deville), while others (e.g., Swain, Markee) are focused on issues of learning.

Larsen-Freeman is concerned with actual language use, including form, meaning and contextual use (that is, pragmatics). It is important, as she points out, to have a large corpus of data in order to understand the when, why, and how of language, and, within the context of this volume, to be able to generalize the data to new contexts. In other words, it is only with large tokens of data in a range of contexts that we can begin to provide answers to the questions of *what* and *why*.

Chapelle, in her paper about vocabulary assessment, makes the important point that we need to consider inferences, that is the interpretation of performance. With regard to generalizabilty, she links the concepts of *inference*, *dependability* (similar to what I referred to above as confidence), and *generalizability*. With specific regard to vocabulary assessment, she brings in a basic assumption that there is similarity across tasks: "[t]he idea is that when performance is observed on one task, it is assumed that this performance is a good representative of what would be displayed, on average, across a large number of similar tasks. . ." (pp. 12–13). This being the case, we infer that task performance on one task generalizes across tasks. If this is the case, then performance is *dependable* across tasks. All of this, as she notes, is dependent on the construct definition that a test is based on.

Both McNamara and Deville & Chalhoub-Deville approach the issue of generalizability from a strict testing framework. McNamara states this succinctly: "Language tests are procedures for generalizing" (p. 2). In fact, the results per se are not of interest. Rather, results are of interest "only as a basis for generalizing beyond the particular performance" (McNamara, p. 10). Thus, the issue that is of interest is: how far can we generalize? For example, Deville and Chalhoub-Deville make clear what language testers aim to do: "We language testers wish to make generalizations – sometimes across test takers, other times across items or tasks, and/or at still other times across occasions or contexts" (p. 2). What is interesting is the notion of "across. . .contexts," which they return to at the end, by calling for "the interaction of language user and context" (p. 19) as the object of study. McNamara picks up on this point by "drawing inferences about the test-taker's ability beyond the immediate testing context" (p. 1). He brings in the concept of sociopolitical issues when talking

about generalizations and makes it clear that any sort of testing or generalization that comes out of it must be cognizant of the consequences of such tests and generalizations.

Thus, within the realm of testing, the idea of the significance of generalizability is unquestioned; rather, the question refers back to the purpose that testing is put to and the context (broadly interpreted) in which testing takes place and which the results often impact.

Swain and Markee both focus on issues of second language acquisition. Markee sees the issue of generalizability as secondary within a conversation analytic perspective. What becomes primary in this tradition is an understanding of how talk is organized. Within a CA context, he demonstrates how a detailed analysis of classroom talk can lead to an understanding of the organization and social interaction of talk in a way that quantitative measures, which would undoubtedly allow greater inferences and generalizability, cannot.

Swain takes the perspective that, in my view, we should be taking when looking at learning. First, she is concerned with the extent to which researchers can generalize from the inferences gained from data collected in "local and situated contexts" (p. 1) (which is clearly related to all data in SLA). Similar to the Polio et al. (2005) study discussed above, she analyzes stimulated recall data from French immersion students. In this and other research (e.g., Swain & Lapkin 1998, 2002, in press), she convincingly shows the role that verbalizations have on cognition. In other words, the recall data actually impacted learner performance. If we return to the notion of generalizability and in particular generalizability in terms of methodology, she argues that methodologies are not neutral. Specifically, with regard to verbal protocols, they are more than a means for eliciting data; they profoundly affect learning.

I return to the issue raised by Larsen-Freeman regarding breadth and abstractness. She specifically ties this into purposes that "applied linguists" deal with. While taking us away from the direct focus of this volume, namely, generalizability, it is important to be clear on "who we are" as researchers. This volume has "Applied Linguistics" in its title and, further, it includes research focused on SLA. It is important, however, that it be understood that not all SLA research has an applied focus (e.g., work focused on SLA from a UG perspective); rather a great deal of SLA work, having no applied focus, is in fact interested in abstractness. Further, for SLA research, it is important, if not crucial, to generalize across languages (both native languages and target languages) and across contexts. One way of doing this is through replication (see Polio & Gass 1997), one way of arriving at a better understanding of the phenomenon of learning. And, perhaps this is the point that Duff is raising when

she discusses controlled and laboratory-like studies – the more controlled, the less generalizable. I would add that this is precisely where replication is necessary. In SLA research, one is often generalizing to models whereas in research with a more applied focus, the focus is on generalizations to populations (see Duff's paper).

With specific regard to generalizability, Duff is particularly clear on this point in stating that with quantitative research it is relatively easy (with appropriate design and sampling procedures) to come up with generalizable results. Duff's point about paradoxes is particularly well-stated: what should we generalize? And how do we know what is generalizable? Quantitative? Qualitative? One case study?

As Applied Linguists, we do live within the real world and are subjected to real-world constraints on data. This is clearly exemplified in Markee's and Duff's papers, who both acknowledge that issues of generalizability are placed in the background of their research paradigms, and perhaps in all or most qualitative research, and in Markee's case the particular context of Conversation Analysis. As Duff says "[g]eneralizability to larger populations, in the traditional positivist sense, or prediction is not the goal" (pp. 8–9) (see also Mackey & Gass 2005, particulary Chapter 6). Duff further points out that the term generalizability itself is a loaded one, belonging more appropriately to a positivist tradition than to a tradition that is more concerned with context-bounded interpretation. But regardless of how we view the notion of generalizability, it is important that we exploit the possible synergy that can be generated by a multiplicity of approaches. So, the tension that is often created between quantitative and qualitative may be a red herring. Rather, it is merely a reflection of our particular research questions, and perhaps the issue is which research area is more interesting. There may theoretically be some mid-point on the continuum where the tension might be real, for example, large-scale research in school settings. Even in this context, however, certain types of research questions truly drive one to one methodology or another. Thus, it may actually not be a question of mid-point tension. Because educational research has to pay attention to student voices and teacher voices and, perhaps more important, the interaction and dynamics of many factors that students face, including the home environment, generalizability may not be as much an issue as is the need to develop an understanding of the various factors that affect children in schools, for example, ESL children in schools. In some arenas, quantitative data are not appropriate (see Crandall 2002). The logistical problems of doing quantitative data collection with immigrant children in a school environment

(e.g., bad data bases, mobility) are many and this may, indeed, be an argument of why qualitative data might be more appropriate for a particular study.

What I want to do in the remainder of this discussion chapter is spin off on the linguistic area of Chapelle's paper, that of acquisition of vocabulary, and in a sense on the content area of the paper by Swain in which she is talking about protocols as an impetus for learning. In particular, I want to problematize the issue of generalizability and deal with the question of learning and knowledge. What is it? And, how do we know when we get there?

We often lose sight of the fact that our definition of learning is dependent on how we view the construct in question, a point that is a focal area of Chapelle's contribution. If we think of a model of learning such as Connectionism, learning involves ever increasing strengthening of connections. Other models of learning take a more traditional view of learning, basing knowledge on correct answers on a test or a certain amount of output. Such concepts as percentage of correct instances in obligatory contexts were prevalent in early research on second language acquisition (e.g., Bailey, Madden, & Krashen 1974; Dulay & Burt 1974; Pica 1984). But it is not uncommon for researchers to use some measure such as 4 out of 6. Talking about generalizability of knowledge or learning cannot take place without a theory of what it means to say that learning has taken place (why would 4/6 be chosen and not 5/6?). Looking at Markee's example, he states that his excerpt illustrates issues about the relationship between second language acquisition and use. Unfortunately, the excerpt didn't go far enough to determine learning, but it appears to be the case that the teacher turns to the class (lines 56–58) to see if someone could define "pretend," but we don't have information about how learning is defined – a necessary prerequisite to saying something about the relationship between acquisition and use. Swain's data are interesting in that she talks about learning moments resulting from verbalization, in particular through verbalization. Perhaps this is what Markee referred to when he mentioned the criticism of self-report data distorting previous behaviour. But Swain's point is an important one: verbalizations are part of the process of cognitive change. They may, in fact, distort previous behaviours, but that is the point. They represent moments of change and, if this is correct, they should "distort" previous behaviours in the sense that they should be different.

In what follows I present an example of some of the difficulties in determining what "knowledge" means and hence in knowing what we are generalizing. As Duff said, and as mentioned above, how do we know what is generalizable? So, with this example, the intent is to problematize some of these

issues. The exchange reported below happened to eight applied linguists and reflects the thoughts of one of them.

Socializing the cost: A learning moment: Or is it?

Location: 8 women, attending a conference in a large Midwestern city, at a French Vietnamese restaurant.

Participants: 6 NSs and 2 NNSs although all the principal players in this episode were NSs of English.
Sharon
Dana
Paula
Mary

Background: Prior discussion about connectionism and differences between representation and processing and what it means to acquire something and how do we measure acquisition (e.g., 90% suppliance in obligatory context?). Discussions included the meaning of learning within a connectionist framework as being strengthening of connections, the possibility that learning constantly takes place if only to involve continued strengthening of associations, and the possibility that learning means ability to generalize to novel contexts. Sharon boldly states that the only learning that will take place at dinner that night would be content learning; she would learn nothing new relating to language.

Immediate context:
Paula was negotiating with the waitperson about the bill. She wanted the wine (2 bottles) to be charged to her and to Mary with the food to be divided amongst the 8 individuals present.

Conversation:

Dana: (overhearing the conversation between Paula and the waitperson): I don't mind socializing the cost.

Sharon: (only partially attending, was struck by the turn of phrase): What?

Dana: I don't mind socializing the cost.

Sharon: (now enamoured by this expression, turns to Mary): You and Paula are picking up the wine tab, but Dana says that she doesn't mind socializing the cost.

Mary: What?

Sharon: Dana says that she doesn't mind socializing the cost of the wine.

What followed was a continuation of this conversation, but now involving a fifth participant. This conversation revolved around the use of this phrase

with the participants attempting to generalize this new phrase to new contexts, for example:

socialize the supervision of theses

but

*socialize the workload.

Dana (the NS of "socializing the cost") and Sharon had similar intuitions about where this phrase could and could not be used. Sharon had by now admitted that she did learn something new about language that evening and had agreed to be evaluated on a posttest in three weeks time.

Following this event, Sharon thought about the conversation and became convinced that she had learned this phrase and had even showed some good knowledge of its limitations (i.e., she had good intuitions); she was worried that she might forget (by now confusing the word with "sharing the cost"). She therefore rehearsed and rehearsed and rehearsed in her mind. So, the question to be asked is: When did learning take place? Was it

– At the moment of hearing the phrase
– During the subsequent discussion and verbalization of the limitations
– During the rehearsal phase
– Or, at a later time when she begins to use it?

Verbalization, as Swain describes it, can definitely be a learning moment, but learning is not often a one-time event; it may require follow-up information (see Gass 1997) which can *inter alia* consist of verbalization, hearing something, thinking about it. I have argued elsewhere that interaction can be a priming device (Gass 1988). It is that moment when attention is drawn to something, but there needs to be follow-up to reinforce or possibly to strengthen associations. I bring up this example and my thinking about it to raise the question in the context of generalizability as to what it is we are generalizing. We need to know when learning takes place to truly understand the concept of generalizability and how it takes place. This is precisely Chapelle's point; we need to understand the construct – knowledge of language. And, finally, we need to know the domain of generalizability.

Note

1. The details of the two groups of teachers go beyond the scope of the paper. However, as a general comment, suffice it to say that, of the preservice teachers, no one had any ESL teaching experience, although 5 of the 11 had tutored non-native speakers of English. In the experienced teacher group, the range of ESL teaching experience was from 4 to 27 years.

References

Bailey, N., Madden, C., & Krashen, S. (1974). Is there a "natural sequence" in adult second language learning? *Language Learning, 24,* 235–243.

Crandall, J. (2002). A delicate balance: The need for qualitative research in an increasingly quantitative environment. Paper presented at *American Association for Applied Linguistics*, April, Salt Lake City.

Dulay, H. & Burt, M. (1974). Natural sequences in child second language acquisition. *Language Learning, 24,* 37–53.

Gass, S. & Mackey, A. (2000). *Stimulated recall in second language research*. Mahwah, NJ: Lawrence Erlbaum.

Gass, S. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics, 9*, 198–217.

Gass, S. (1997). *Input, interaction and the development of second languages*. Mahwah, NJ: Lawrence Erlbaum.

Mackey, A. & Gass, S. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

Mackey, A., Polio, C., & McDonough, K. (2004). The relationship between experience, education, and teachers' use of incidental focus-on-form techniques. *Language Teaching Research, 8,* 301–327.

Pica, T. (1984). Methods of morpheme quantification: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition, 6,* 69–78.

Polio, C. & Gass, S. (1997). Replication and reporting: A commentary. *Studies in Second Language Acquisition*, *19*, 499–508.

Polio, C., Gass, S., & Chapin, L. (2005). Preservice and Experienced Teachers' Perceptions of Feedback during Interaction. Paper presented at *Voice and Vision in Language Teacher Education Conference*, University of Minnesota, June.

Swain, M. & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *Modern Language Journal, 82*, 320–337.

Swain, M. & Lapkin, S. (2002). Talking it through: Two French immersion learners' response to reformulation. *International Journal of Educational Research, 37*, 285–304.

Swain, M. & Lapkin, S. (in press). "Oh, I get it now!" From production to comprehension in second language learning. In D. M. Brinton & O. Kagan (Eds.), *Heritage language acquisition: A new field emerging*. Mahwah, NJ: Lawrence Erlbaum.

Negotiating methodological rich points in applied linguistics research

An ethnographer's view*

Nancy H. Hornberger
University of Pennsylvania

"Methodological rich points" are those points that make salient the pressures and tensions between the practice of research and the changing scientific and social world in which researchers work – points where our assumptions about the way research works and the conceptual tools we have for doing research are inadequate to understand the worlds we are researching. In this paper, I highlight some methodological rich points around issues of inference and generalizability in applied linguistics research, drawing on the papers in this volume and on my own and others' ethnographic applied linguistics research. At the same time, I seek to reframe these issues in the context of more basic questions of research methodology and ethics.

Introduction

Applied linguists research language learning and teaching in a range of settings including classrooms, communities, test and experimental situations, using a variety of methods, tools, and strategies, in order to understand, inform policy, and transform language teaching/learning. This is, at its most basic, the applied linguistics research endeavor. Within that endeavor, there are diverse ways applied linguists might approach issues of inference and generalizability, and the authors in this volume provide a substantial and thoughtful sample, exploring limits and possibilities of these principles in advancing applied linguistics research within their particular specializations.

For me as an ethnographer of language and education, these papers provoke reflection on what I have come to think of as "methodological rich points," points that make salient the pressures and tensions between the practice of re-

search and the changing scientific and social world in which researchers work.[1] In other words, methodological rich points are those times when researchers learn that their assumptions about the way research works and the conceptual tools they have for doing research are inadequate to understand the worlds they are researching. When we pay attention to those points and adjust our research practices accordingly, they become key opportunities to advance our research and our understandings.

I have borrowed and adapted the term methodological rich points from ethnographer Michael Agar's notion of "rich points" as those times in ethnographic research when something happens that the ethnographer doesn't understand; those times when "an ethnographer learns that his or her assumptions about how the world works, usually implicit and out of awareness, are inadequate to understand something that had happened" [in the corner of the world he or she is encountering] (1996:31). Agar discusses rich points as one of three important pieces of ethnography: participant observation makes the research possible, rich points are the data you focus on, and coherence is the guiding assumption by which you seek out a frame within which the rich points make sense (1996:32). Rich points, then, are points of experience that make salient the differences between the ethnographer's world and the world the ethnographer sets out to describe. Methodological rich points are, by extension, points of research experience that make salient the differences between the researcher's perspective and mode of research and the world the researcher sets out to describe.

The authors herein provide insight into just such methodological rich points around issues of inference and generalizability in applied linguistics research. In what follows, I try to highlight these, drawing on the papers in this volume and on my own and others' ethnographic applied linguistics research. At the same time, I seek to reframe these issues in the context of more basic questions of research methodology and ethics. As a sociolinguist, I borrow as organizing rubric to do this the paradigmatic heuristic for sociolinguistic analysis first offered by Fishman (1971:219) and here adapted to applied linguistics research to ask: who researches whom and what, where, how, and why?

Who researches whom in applied linguistics?

At the most basic level, applied linguists research language learners and users – teachers, students, community members, policy makers. The papers here are no exception – the research described is that of applied linguists researching

students, teachers, and language learners. There is not a lot of problematizing of the researcher/researched relationship, beyond on the one hand a concern with finding ways to involve participants in the research to enhance findings, and on the other a concern with paying attention to the social consequences of testing practices for those tested. The former is exemplified by involving participants in member checks (Duff) or stimulated recall (Swain) as ways of deepening the researchers' understanding and interpretation of observed behaviors. The latter is explicitly addressed as an aspect of validity in Messick's framework as discussed by McNamara.

For Duff, the use of member checks, whereby the researcher consults with participants during analysis and write-up of the findings, is one of several ways of attending to establishing the credibility of research findings and generalizations (along with other strategies such as sufficient amount and diversity of data, consideration of counter-examples, triangulation not only of data and methods but also of theory and researchers). Beyond consulting participants for the sake of credibility, there is, particularly in ethnographic research, another set of participant-related concerns I want to highlight here as a methodological rich point – these are the concerns around questions of collaboration, authority and representation.

Authority refers to the researcher's authority over the interpretation of the data – the right to claim that he or she has 'got it right' in reporting findings. On what basis does the researcher have (or not) authority to speak for the participants (the researched)? This issue is closely linked to that of collaboration. "Ethnographic research is collaborative ... It's always been that way... What the new ethnography calls for is attention to the way collaborative work leads to the results" (Agar 1996: 16). Agar goes on to note that the authority issue also puts the spotlight on the ethnographer and the question of who studies whom (1996: 17), leading to questions about who is self and who is other, and even what is emic and what is etic (Agar 1996: 21).

In applied linguistics research, this methodological rich point has been forcefully and articulately raised in terms of the slogan "research on, for, and with subjects," put forward by Cameron, Rampton and colleagues (Cameron et al. 1992). After first discussing issues of power and of positivist, relativist, and realist paradigms of research, the authors introduce a distinction between an ethics-based approach (research on subjects), which seeks to balance the needs of a discipline in pursuit of knowledge with the interests of the people on whom the research is conducted; an advocacy-based approach (research on and for subjects), which despite its commitment to participants nevertheless still tends toward a positivist notion that there is one true account; and an empowerment-

oriented approach (research on, for, and with subjects), which uses interactive, dialogic methods and seeks to take into account the subjects' research agenda, involve them in feedback and sharing of knowledge, consider representation and control in the reporting of findings, and take seriously the policy-making implications of the research. The authors clearly advocate the last approach and offer examples of attempts to implement it in their own research.

Educational ethnographer Reba Page speaks of a crisis in representation in qualitative research. She writes that increasing recognition of limits to "the qualitative claim that researchers could document and explain, fully and accurately, another's life-world as it is" (Page 2000: 5) presents both an aesthetic challenge centering on how knowledge is represented in texts and a political challenge centering around whose representations are the ones put forward. The political challenge has given rise to new interdisciplinary alignments, fieldwork relations, and advocacy stances (Page 2000: 6–7). Similarly, in response to the aesthetic challenge, scholars have "experimented with modes of reproduction that [give] more prominence to their own meaning-making, the artfulness of accounts, and the diverse 'voices' and alternate views of informants" in the form of dialogic scripts, collaborative authorship, autobiographical ethnographies, and even novels (Page 2000: 6).

Paying more attention to issues of authority, collaboration, and representation in applied linguistics research may take a number of forms – it may be about working with multiple members of a research team; it may also be about relationships between researcher and researched; and may range from consultative to fully participatory relationships. It may be about collecting and analysing data; it may also be about writing up and reporting findings. It is without doubt about incorporating multiple voices in the research process and producing multi-voiced texts.

What do we research?

A succinct (and partly oversimplified) way of stating what it is that applied linguist researchers concern ourselves with is that we research language teaching and learning, as well as *the role of* language *in* learning and teaching (cf. Hornberger 2001 on educational linguistics). In the papers collected here we find these concerns instantiated in investigations of: language learning and instructional practices in classrooms at elementary, secondary, and college levels (Duff), language acquisition and use in classroom interaction (Markee), the relation between language and thought (Swain), the ways people mobilize gram-

matical resources to accomplish communicative and social purposes (Larsen-Freeman), the acquisition of vocabulary in a second language (Chapelle), and methods of assessing what a language learner knows and can do (McNamara; Deville & Chalhoub-Deville).

Striking for an ethnographer in reading these essays is the extent to which a qualitative concern for context and contextualization comes up in relation to the content (the "what") of the authors' research and its validity, reliability, or dependability, even in research which does not necessarily use qualitative methods. McNamara discusses at length the four quadrants of Messick's validity framework for test constructs, of which three are about contextual uses, values, and consequences attached to the test; he also draws attention to the growing body of work using discourse-based approaches to the validation of oral language assessment. Larsen-Freeman appeals to contextual analysis as a research methodology for accounting for form, meaning, and use of linguistic forms – in this case, context refers primarily to the discourse context in which naturally observed spoken and written language occurs. Swain argues that verbal protocols are not simply a means of reporting what one is thinking, but rather actually enable changes in thinking; they are therefore not a neutral means of collecting data (i.e. content), but in fact are part of the treatment (i.e. context). Markee demonstrates through detailed conversation analysis of a stretch of classroom interaction that the social relationship between the teacher and learner (i.e. language-learning context) had to be repaired before further language-learning oriented talk (i.e. language-learning content) could occur. Meanwhile, Duff, writing explicitly as a qualitative researcher, includes contextualization and ecological validity of tasks among other means of addressing internal validity in qualitative research; and goes on to provide examples from her own ethnographic classroom research in Hungary of how contextualization – specifically, attention to sociopolitical and historical context, the participants and their interests, the tasks or instructional practices used and the participants' understandings of these – is crucial for establishing internal validity of her findings. Duff emphasizes the need to consider context from the point of view of how "sociopolitical structure not only influences and mirrors, but also is constituted in" language learning and teaching events and interactions in everyday classrooms.

These various insights point to another methodological rich point in applied linguistics research – namely the recognition that knowledge, and specifically language learning, is co-constructed in contexts of social interaction. McNamara takes note of the growing critique of individualistic models of proficiency, in favor of work that stresses the co-construction of performance

(raising questions around, for example, what exactly is being assessed in an oral proficiency interview). Markee offers a social constructivist critique of second language acquisition (SLA) research on social interaction. Duff recalls psychologist Lee Cronbach's early (1982) paradigm-shifting acknowledgement that "human action is constructed, not caused." Swain adopts as premise that language, and specifically verbalization, is constitutively involved in thinking. Larsen-Freeman is interested in how people use language (specifically grammatical resources) to, among other things, manage interpersonal relationships and position themselves socio-politically. These observations rest on social constructionism, the "view that instead of being the product of forces that actors neither control nor comprehend, human reality is extensively reproduced and *created anew* in the socially and historically specific activities of everyday life" (Rampton 2000: 10; citing Giddens 1976, 1984).

For sociolinguistic ethnographers and linguistic anthropologists of education, social and cultural context has always been the bedrock of research method, evident in such long established strands of work as the ethnography of communication – documenting and comparing ways of speaking (Hymes 1964, 1968; Philips 1983; Heath 1983), interactional sociolinguistics – revealing the multiple linguistic means by which we embed social meanings in interaction (Gumperz 1982), and microethnography – demonstrating the importance of situationally emergent social identity and co-membership (Erickson & Shultz 1982) (see Hornberger 1995 for a review of these three sociolinguistic approaches to school ethnography). More recently, scholars working in these traditions have begun to frame their research with more explicit attention to social constructionism, documenting patterns of language use and social relations in multilingual classrooms and communities, and exploring dimensions of discourse that maintain the status quo in societal power relations (e.g. Martin-Jones & Jones 2000; Heller & Martin-Jones 2001; Hornberger 2003; Wortham & Rymes 2003; Creese & Martin 2003; Arkoudis & Creese 2005; McCarty 2005). There is also increasingly explicit attention in this work to the heterogeneous, multilingual, multicultural, multiliterate classroom and community contexts in which language learning, teaching, and use take place. This we take up next.

Where do we research?

Classrooms at all levels, community sites of formal and non-formal education, testing and (semi-) experimental situations are the usual venues for applied

linguistics research, and the papers herein are no exception. In keeping with the volume's theme of generalizability, the typicality or representativeness of sites or cases comes up in several of the papers, usually with the cautionary acknowledgement that typicality is quite elusive (and perhaps in some cases too easily assumed). Larsen-Freeman points out that intuitional data on sentence grammaticality must be complemented by other data (observational and elicited), since "intuitions can be unreliable or undependable when looking for typical patterns because humans tend to notice the unusual more than the typical." Duff reminds us that typicality may not always be the aim in selecting sites or cases, anyway; atypical, unique, resilient, extreme, or even pathological cases or instances may be purposely sought out for the potential insight they offer – these are so-called intrinsic cases, in Stake's (2000) terms, as distinct from instrumental cases examined explicitly in relation to a generalization.

Markee offers a two-pronged argument about research settings and typicality in discussing what he calls the domain problem, following Schegloff (1993). On the one hand, he acknowledges that Conversational Analysis (CA) originally focused on the study of ordinary conversation and has only more recently taken up the analysis of institutional talk (e.g. news, medical, courtroom, or classroom talk), and that one can't necessarily generalize from the former speech exchange system to the latter. On the other, he argues that the predominantly experimental tradition in second language acquisition research on negotiated interaction has not adequately studied the interactional speech exchange system (i.e. domain) in which the negotiation occurs. He suggests that CA work on classroom interaction may provide a means to address this mutual lack and a qualitative basis on which to build a quantified search for generalizations.

Apart from typicality of settings, the papers also allude to the complexity of any particular research setting. Indeed, when Duff suggests that one means of enhancing generalizability of cases or settings is to conduct multi-site or multiple-case studies, she acknowledges that this nevertheless brings with it a concomitant reduction of possibilities for in-depth description and contextualization of the kind that would do justice to the complexity of any one site or case.

For applied linguists, one very clear aspect of research setting complexity – one which constitutes another methodological rich point for applied linguistics research – is the increasingly diverse range of settings where we do research, along with the increasingly heterogeneous, multilingual nature of those settings. So that, in addition to research in classrooms or language learning or testing settings, we also have (comparative) ethnographic studies

of language, literacy, and education in school-and-community settings (e.g. Heath 1983; Hornberger 1988; Delgado-Gaitan 1990; McLaughlin 1992), in out of school settings such as adult literacy programs, workplaces, religious settings (e.g. Heath & McLaughlin 1993; Spener 1994; Knobel 1999; Hull & Schultz 2002), in bilingual and multilingual classroom settings around the world (e.g. Heller & Martin-Jones 2001; Creese & Martin 2003), in language education professional development and practice settings (e.g. Henning 2000; Pérez et al. 2003; Brutt-Griffler & Varghese 2004; Hawkins 2004; Arkoudis & Creese 2005), and in language education policy-making settings and activities (e.g. Tollefson 1995, 2002; Freeman 2004; Johnson 2004; Tollefson & Tsui 2004; Canagarajah 2005b).

For applied linguist ethnographers, the crux of the methodological rich point here is the changing nature – or perhaps more accurately, the deepening understanding – of the concept of speech community as research setting for the ethnographic study of language use and language learning. Defined in sociolinguistic work of the 1960s as a community whose members share at least one language variety and the norms for its use (Hymes 1972: 54; Fishman 1971: 232), the underlying assumption in the concept is not that there is uniformity of communicative resources and practices within a speech community, but rather that there is a patterned diversity of those resources and practices: as Hymes often repeated, it is "not replication of uniformity but organization of diversity" (Hymes citing Wallace 1961). The task of the ethnography of a speech community is to "Take as context a community, investigating its communicative habits as a whole, so that any given use of channel and code takes its place as but part of the resources upon which the members of the community draw" (Hymes 1964: 3).

In response to the rise of post-modernity and social constructionism, Rampton (2000) tells us, analysis of the speech community has moved on the one hand toward "investigation of 'community' as itself a semiotic sign and ideological product" and on the other toward "close-up analysis of face-to-face interaction in relatively consolidated social relationships", sometimes termed "communities of practice (e.g. unions, trades, boards of directors, marriages, bowling teams, classrooms)", i.e. "a range of social relationships of varying duration" (Rampton 2000: 10, 12). Since the 1990s, there is a shift from an exclusive focus on speech communities (and descriptions of deficit, difference or domination across them) to also focus on communities of practice and communicative practices, yielding "fine-grained and complex account[s] of imposition, collusion and struggle" (Rampton 2000: 12), where randomness and disorder are more important than system and coherence, and anoma-

lous social difference is treated as central rather than peripheral (Rampton 2000: 9, 18). These tendencies are evident in ethnographic research in all the variety of speech communities mentioned above, from classroom to school to out-of-school and beyond.

Indeed, this methodological rich point is not only about the increasingly heterogeneous nature of any one research setting, but also about the increasingly diverse range of settings in which applied linguistics research takes place – from face-to-face interactions at the micro-level to policy discourses and globalizing forces at the macro-level. This is particularly true of ethnographic and sociolinguistic studies, which increasingly turn their attention toward the discourses of language planning and policy as well as those of classroom and community interaction, encompassing both the global and the local. There is growing recognition that language planning and policy-making happens as much at the micro-level of the classroom as it does at the macro-level of government (Ricento & Hornberger 1996; Ricento 2006); acknowledgement of the tensions in language in education policies and practices, especially in post-colonial contexts undergoing simultaneous and contradictory processes of decolonization and globalization (Lin & Martin 2005); and movement toward a more localized orientation that takes seriously the tensions, ambiguities, and paradoxes of language allegiances and sociolinguistic identities in order to construct policies from the ground up (Canagarajah 2005a; see also Hornberger 1996). Ecological approaches, in particular, have been proposed as a way to do this (Hornberger 2003; Canagarajah 2005a, 2005b).

How do we collect, analyze, and interpret data?

As applied linguists, our primary data are bits or stretches of spoken or written language, collected primarily by observation, recording, elicitation, and testing; analyzed usually in some way for form, function, and meaning; and interpreted within a variety of conceptual frameworks ranging from highly specified to more loosely configured. Among the methods of data collection discussed or exemplified in this volume are transcription of recorded ordinary conversation/institutional talk (Markee); ethnographic participant observation (Duff); intuition, observation and elicitation (Larsen-Freeman); collaborative dialogue, verbal protocols including concurrent think-alouds and retrospective introspection or stimulated recall, and videotaped interaction in the classroom (Swain); and assessment or testing (McNamara; Chapelle; Deville & Chalhoub-Deville).

All the authors define and discuss how they approach inference, as integral to processes of data analysis and interpretation. Whether it be from people's observable behavior to their underlying knowledge systems (Duff), from language performance to forms, meanings, functions (Larsen-Freeman), from observed test performance to performance under non-test conditions (McNamara), or from observed performance to what the performance means (Chapelle), inference is always about the logical connection the researcher draws from research evidence to claims about that evidence – and also reflexively back on the inferential construct itself. So McNamara tells us that test data can be used to validate inferences we draw based on constructs, but also to validate the constructs themselves; while Chapelle distinguishes in vocabulary acquisition research between construct inference (from performance to vocabulary construct) and theory inference (from vocabulary construct to a theory of vocabulary knowledge). Markee argues that CA work on classroom interaction could provide the missing logical link from the accumulated evidence provided by experimental SLA research on negotiated interaction to theoretical claims about language learning.

The authors also problematize inference in a variety of ways. Larsen-Freeman reminds us that inferences are always provisional and partial, subject to refutation as other data are considered; but she also points out that more data do not necessarily lead to better inferences – after all, a large corpus may increase dependability of observed patterns, but may also reveal more variations within the patterns; and it remains true that insightful inferences have often been drawn from one or only a few instances; furthermore, since language is always changing, it's a moving target in any case. Markee muses on the greatest inferential puzzle of SLA, namely the question of "how psycholinguistic questions of language learning intersect with sociolinguistic aspects of language use"; and he also raises the "significance" problem (following Schegloff), pointing out that conclusions drawn about a great number of instances of interactional repair are not useful if the categories of repair are ambiguous and decontextualized.

Other authors demonstrate that different inferences can be drawn from the same data, based on one's theoretical approach or conceptual framework and concomitant construct. Chapelle reminds us that construct definition for vocabulary assessment may not look exactly the same when undertaken from trait, behaviorist, or interactionist perspectives. Swain demonstrates that different inferences may be drawn from verbal protocols by information processing theorists (who see language primarily as communication of thought) and

sociocultural theory of mind theorists (who see language as crucially implicated in human thinking).

The issues these authors highlight – sufficiency of data as a basis for inference and the inferential relation between theory and data – constitute methodological rich points which are if anything even more central in ethnographic applied linguistics research, by definition interpretive. Interpretation is after all a kind of inference, a search for patterns and understandings (Duff). Interpretation, and induction, both terms used more often than inference in ethnographic research, explicitly highlight the subjective involvement of the ethnographer in mediating between theory and data, crucial to achieving two of ethnography's defining characteristics – a holistic and an emic view. Emic in that the ethnographer attempts to infer the "native" point of view: to describe the culture, cultural situation, or cultural event as its members understand it and participate in it (i.e. as they make sense of it). Holistic in that the ethnographer seeks to create a whole picture, one that leaves nothing unaccounted for and that reveals the interrelatedness of all the component parts (Hornberger 1992: 186, 1994: 688).

The emic/etic distinction so often invoked in ethnographic research was first proposed by Pike (1954), in direct parallel to the phonemic/phonetic distinction in phonology. In the study of human behavior, the etic standpoint is one situated outside the system studied, in which units and classifications are determined on the basis of existing knowledge of similar systems, and against which the particular system is measured; while the emic standpoint is one situated inside the particular system studied, which views the system as an integrated whole, and in which units and classifications are determined during and not before analysis, and are discovered and not created by the researcher. Both standpoints are necessary and it is the movement back and forth between them that takes our understanding forward. Hymes speaks of Pike's three moments, etic-1, emic, etic-2, in terms of a "dialectic in which theoretical frameworks are employed to describe and discover systems, and such discoveries in turn change the frameworks" (Hymes 1990: 421). This dialectic movement from theory to data and back again is essential to the process of ethnographic interpretation, and it is the ethnographer who provides the inferential link.

In a recent essay on the development of conceptual categories in ethnographic research, Sipe and Ghiso emphasize the paradoxical nature of the interpretive process wherein "theoretical frameworks are essential to structuring a study and interpreting data, yet the more perspectives we read about, the greater the danger of overdetermining conceptual categories and the ways in which we see the data" (Sipe & Ghiso 2004: 473). Demonstrating a process in

which "induction and deduction are in constant dialogue" (Erickson 1986: 21), Sipe and Ghiso provide a detailed example of a breakthrough in Sipe's development of categories for his classroom data that came precisely from his reading Bakhtin at the time. In a commentary on their essay, Erickson underlines this point, noting that if Sipe had been reading someone else, e.g. Fish, Foucault, or Habermas, the analysis might have gone in a different direction (Erickson 2004: 489).

Part of what drove Sipe to look for a further category in the first place was the existence of data that didn't fit the categories he had worked out up to that point – outlier data that he became increasingly uncomfortable categorizing as simply "off-task" (Sipe & Ghiso 2004: 480). It was a question of sufficiency not so much in the *amount* of data as the *kinds* of data that posed an inferential challenge for Sipe. Erickson comments on this, too, noting that whereas *quantitas* is always first about "what amounts?," *qualitas* is about "what kinds?" (Erickson 2004: 487). Grappling with the data that didn't fit, the discrepant cases, Sipe achieved an interpretive breakthrough – and Erickson reinforces this point, emphasizing that Sipe's example demonstrates that neither ethnographic data themselves nor interpretive themes and patterns simply emerge, but rather must be found by the researcher (Erickson 2004: 486).

Erickson applauds Sipe & Ghiso's demystification of ethnographic data construction and analysis and takes the demystifying process one step further by considering alternative approaches to the "exhaustive analysis of qualitative data," contrasting Sipe's bottom up approach with a top down approach that would "parse analytically from whole to part and then down again and again, successively identifying subsequent next levels and their constituents at that level of contrast [rather than] start by trying to identify parts first and then work up analytically from there" (Erickson 2004: 491). He prefers the top down approach himself in part because he thinks that is what social actors do, and in part because it invites "parsing all the way down on both sides of [the] analytic divide" (Erickson 2004: 491). Whether bottom up or top down, the quest is for holism. It is, ultimately, the holistic and emic quality of the ethnographer's account that determines its validity and generalizability. In a similar move toward demystifying data construction and analysis, Chapelle argues for the centrality of construct definition in vocabulary acquisition research and suggests that recent explicit discussion of vocabulary knowledge as a construct seems to mark progress toward a more generalized theory of L2 vocabulary acquisition. These allusions to generalizability bring us to the last part of our heuristic question.

Why do we research?

The goal of applied linguistics research is, at its most fundamental, to understand and inform the teaching and learning of language. Through their research, applied linguists seek to understand processes, inform policies, and transform practices of language learning and teaching. In this vein, Larsen-Freeman identifies three purposes in her work: to inform the identification of the language acquisition/language learning challenge; to better understand processes contributing to or interfering with meeting that challenge; and to adopt pedagogical strategies, design materials, and educate teachers, based on understandings of those challenges and processes. Markee seeks to explicate the complex phenomenon of how and why second languages are learned, Chapelle to evaluate both materials and tests used in L2 vocabulary learning and teaching, and Swain to shed light on the role of verbalization in language learning and in research on language learning. McNamara and Deville & Chalhoub-Deville look at issues around the construction, validation, and use of tests in language learning and teaching.

In all of this and for all the authors herein, generalizability becomes an important consideration. McNamara points out that, of all the fields of applied linguistics, language testing is probably where questions of inference and generalizability are of most concern, since language tests are in essence procedures for generalizing. Yet it is also true that in every area of applied linguistics research, the overall aim of generalizability – i.e. "establish[ing] the relevance, significance and external validity of findings for situations or people beyond the immediate research project" (Duff) – holds true. A number of distinctions and related terms are discussed in the papers. Duff, following Firestone (1993), adopts the term analytic generalizability to refer specifically to generalization at an abstract or theoretical level, as distinct from generalization at the level of cases, populations or sites. Chapelle differentiates between a generalization inference (from performance on one task to performance on other similar tasks) and an extrapolation inference that generalizes from performance on the task (or tasks) to performance beyond the task. She notes that generalization inference is the same as dependability; and Deville & Chalhoub-Deville likewise use generalizability, dependability, and reliability interchangeably.

Concerns about generalizability raised by the authors revolve in particular around notions of variability and abstraction. Deville & Chalhoub-Deville emphasize that variability is critical to any discussion of reliability (and generalizability); Duff warns that generalizability may be inadvertently reduced by key sociocultural contextual variables, regardless of claims made by re-

searchers. Larsen-Freeman tellingly depicts that the price for increased generalizability is increased abstraction, to the point that one loses any sense of the particularity of the case; and Swain problematizes the extent to which we can generalize from inferences drawn from local and situated contexts. In addition to Schegloff's domain and significance problems mentioned in earlier sections, Markee discusses Schegloff's denominator and numerator problems, arguing that the quantified search for generalizability is premature and possibly misleading when it is not preceded by adequate specification of the analytically relevant denominator (e.g. repairs per task type) and numerator (e.g. absence as well as presence, rarity as well as frequency of repairs). With regard to procedures for generalizing from tests, McNamara addresses challenges arising from social science critiques of an overly abstract positivism, in relation on the one hand to the particular values implied within test constructs and on the other to the particular social and political uses made of test scores.

These concerns around variability and abstraction constitute another methodological rich point in applied linguistics research, one that ethnographers tend to grapple with in terms of transferability and particularity. Transferability, as Duff points out, assigns responsibility to readers to determine whether findings apply to another context; variability across contexts is taken for granted, but if the ethnographer provides enough rich and detailed description of one local context, it should be possible for the reader familiar with another local context to sort out what findings might or might not transfer. In that regard, the greater the particularity of description and interpretation (and the less the abstraction), the more likely it is that a reader will be able to determine whether these particular findings apply to another context. In this sense, Duff suggests that in qualitative research, the goal is not generalization or prediction but rather a search for particularity – what Geertz famously called "thick description" (Geertz 1973).

Final reflections

Across all the methodological rich points these essays highlight for me, there is a common thread of recognition of the existence of multiple possible actors, multiple possible trajectories, and multiple possible truths (Duff) in applied linguistics research. This is all to the good, not because they should or could all be combined in some grand "mixed methods" research project; I am not much for mixed methods approaches anyway, partly because I don't think they're particularly new and partly because in my experience there is usually one dom-

inant method subsuming the others. Rather, the value of multiple research actors, trajectories, and truths is that they all offer different versions of and perspectives on the vast language learning and teaching enterprise. To the degree that these multiple approaches share common methodological rich points such as those highlighted here, they are all potentially enriched by dialogue across and within the applied linguistics field.

Notably, along with inference and generalizability, these authors have also talked about credibility and contextualization in relation to research questions, collaboration and representation in relation to research participants, heterogeneity and co-construction in relation to research methods, demystification and ecology in relation to analytical approaches. Underpinning all these are critical concerns that go beyond inferring or generalizing findings to transforming the realities they describe; there is increasingly explicit attention to power and inequality and the role of research and of the researcher in interrogating those. As Agar puts it in relation to critical ethnography, "you look at local context and meaning, just like we always have, but then you ask, *why* are things this way? What power, what interests, wrap this local world so tight that it feels like the natural order of things to its inhabitants?" (Agar 1996: 26). Or to paraphrase Pennycook on discourse analysis in applied linguistics research, the critical question becomes not only what language means and how that meaning is constructed across sentences, but also why those particular meanings out of all possible available meanings are expressed at that particular moment in time and place (Pennycook 1994: 116). As research increasingly locates communicative practices as parts of larger systems of social inequality (Gal 1989: 347), it becomes natural to ask what we, as researchers, can do about transforming those practices and those inequalities.

As applied linguists set about that task in our multiple and varied research endeavors, there are two more methodological rich points which to me seem basic – humility and respect. Humility before the rich diversity of language learning and teaching practices and contexts we have the privilege to observe and seek to understand, and respect for the language teachers, learners, and users, both individuals and communities, who untiringly and insightfully ply their language and pedagogical knowledge and skills, day in and day out the world over.

I close with an example of ethnographic work in language and education that epitomizes many of the methodological rich points highlighted above. Pippa Stein recounts experiences with two projects she has worked on with pre-service and in-service language teachers in Johannesburg, both of which encourage students' use of a range of representational resources in their mean-

ing making, including the linguistic mode in its written and spoken forms, the visual, the gestural, the sound, and the multimodal performance (Kress & Van Leeuwen 1996; Kress 1997). A reflective practitioner, she is exploring "pedagogies in which the existing values attached to representational resources are reconfigured to take into account a broader notion of semiotic resources" (Stein 2004: 37), specifically with the goal of ascribing equal value to resources brought by historically advantaged and historically disadvantaged students. Both the Performing the Literacy Archive Project and the Photographing Literacy Practices Project focus on literacy because "issues of literacy and access to it are at the heart of educational success in South African schools" (2004: 37) but in them the students explore "the use of multiple semiotic modes in the making of meaning" (2004: 37). Drawing on her reflections and on written and video documentation of the students' work over the several years she has done these projects with language teachers, she shows how these pedagogies "work with what students bring (their existing resources for representation) and acknowledge what [historically disadvantaged] students have lost" (2004: 50). As she puts it, it is "the saying of the unsayable, that which has been silenced through loss, anger or dread, which enables students to re-articulate their relationships to their pasts. Through this process of articulation, a new energy is produced that takes people forward. I call this process of articulation and recovery re-sourcing resources" (2004: 39). Stein's reflective practitioner account is about transforming language teaching and learning; that, I believe, is what applied linguistics research is most fundamentally about.

Notes

* This paper is based (very loosely) on my keynote talk at the 2000 Conference on Qualitative Research in Education at (then) Rand Afrikaans University in Johannesburg, South Africa. I would like to express my appreciation to Elizabeth Henning, who invited me and thereby got me started thinking on many of these issues, and whose own ethnographic work in teacher education (e.g. Henning 2000) beautifully expresses humility of stance and respect for persons in her research context.

1. Methodological rich points are akin to Eisenhart's (2001) muddles in educational ethnography. She discusses three, which are also reflected here to some degree: (1) the meaning of culture in postmodern times, (2) the increasing popularity of ethnographic research across disciplines, along with the backlash against it, and (3) the ethics of representation.

References

Agar, M. (1996). Ethnography reconstructed: The professional stranger at fifteen. In M. Agar, *The professional stranger* (pp. 1–51). New York, NY: Academic Press.

Arkoudis, S. & Creese, A. (Eds.). (2005). Teacher teacher talk in multilingual contexts. *International Journal of Bilingual Education and Bilingualism*. Special issue.

Brutt-Griffler, J. & Varghese, M. M. (Eds.). (2004). *Bilingualism and language pedagogy.* Clevedon, UK: Multilingual Matters.

Cameron, D., Frazer, E. et al. (1992). *Researching language: Issues of power and method*. London: Routledge.

Canagarajah, A. S. (2005a). Accommodating tensions in language-in-education policies: An Afterword. In A. Lin & P. Martin (Eds.), *Decolonisation, globalisation: Language-in-education policy and practice.* Clevedon, UK: Multilingual Matters.

Canagarajah, A. S. (Ed.). (2005b). *Reclaiming the local in language policy and practice*. Mahwah, NJ: Lawrence Erlbaum.

Creese, A. & Martin, P. (Eds.). (2003). *Multilingual classroom ecologies: Inter-relationships, interactions and ideologies*. Clevedon, UK: Multilingual Matters.

Cronbach, Lee J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Delgado-Gaitan, C. (1990). *Literacy for empowerment: The role of parents in children's education*. London: Falmer Press.

Eisenhart, M. (2001). Educational ethnography past, present, and future: Ideas to think with. *Educational Researcher, 30*(8), 16–27.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 119–161). New York: Macmillan.

Erickson, F. (2004). Demystifying data construction and analysis. *Anthropology and Education Quarterly, 35*(4), 486–493.

Erickson, F. & Shultz, J. (1982). *Counselor as gatekeeper: Social interaction in interviews*. New York, NY: Academic Press.

Firestone, W. A. (1993). Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher, 22*, 16–23.

Fishman, J. A. (1971). The sociology of language: An interdisciplinary social science approach to language in society. In J. A. Fishman (Ed.), *Advances in the sociology of language* (pp. 217–237). The Hague: Mouton.

Freeman, R. D. (2004). *Building on community bilingualism*. Philadelphia, PA: Caslon Publishing.

Gal, S. (1989). Language and political economy. *Annual Review of Anthropology, 18*, 345–367.

Geertz, C. (1973). *The interpretation of cultures: Selected essays*. New York, NY: Basic Books.

Giddens, A. (1976). *New rules of sociological method*. London: Hutchinson.

Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Berkeley, CA: University of California Press.

Gumperz, J. J. (1982). *Discourse strategies*. Cambridge: CUP.

Hawkins, M. R. (Ed.). (2004). *Language learning and teacher education: A sociocultural approach*. Clevedon, UK: Multilingual Matters.

Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms*. New York, NY: CUP.
Heath, S. B. (2000). Linguistics in the study of language in education. In B. M. Brizuela, J. P. Stewart, et al. (Eds.), *Acts of inquiry in qualitative research* (pp. 27–36). Cambridge, MA: Harvard Educational Review.
Heath, S. B. & McLaughlin, M. (Eds.). (1993). *Identity and inner-city youth: Beyond ethnicity and gender*. New York, NY: Teachers College Press.
Heller, M. & Martin-Jones, M. (Eds.). (2001). *Voices of authority: Education and linguistic difference*. Norwood, NJ: Ablex.
Henning, E. (2000). Walking with "barefoot" teachers: An ethnographically fashioned casebook. *Teaching and Teacher Education, 16*, 3–20.
Hornberger, N. H. (1988). *Bilingual education and language maintenance: A southern Peruvian Quechua case*. Berlin: Mouton.
Hornberger, N. H. (1992). Presenting a holistic and an emic view: The Literacy in Two Languages project. *Anthropology and Education Quarterly, 23*, 160–165.
Hornberger, N. H. (1994). Ethnography. *TESOL Quarterly, 28*(4), 688–690.
Hornberger, N. H. (1995). Ethnography in linguistic perspective: Understanding school processes. *Language and Education, 9*(4), 233–248.
Hornberger, N. H. (Ed.). (1996). *Indigenous literacies in the Americas: Language planning from the bottom up*. Berlin: Mouton.
Hornberger, N. H. (2001). Educational linguistics as a field: A view from Penn's program on the occasion of its 25th anniversary. *Working Papers in Educational Linguistics, 17*(1/2), 1–26.
Hornberger, N. H. (Ed.). (2003). *Continua of biliteracy: An ecological framework for educational policy, research and practice in multilingual settings*. Clevedon, UK: Multilingual Matters.
Hull, G. & Schultz, K. (Eds.). (2002). *School's out! Bridging out-of-school literacies with classroom practice*. New York, NY: Teachers College Press.
Hymes, D. H. (1964). Introduction: Toward ethnographies of communication. *American Anthropologist, 66*(6), 1–34.
Hymes, D. H. (1968). The ethnography of speaking. In J. A. Fishman (Ed.), *Readings in the sociology of language* (pp. 99–138). The Hague: Mouton.
Hymes, D. H. (1972). Models of the interaction of language and social life. In J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics: The ethnography of communication* (pp. 35–71). New York, NY: Holt, Rinehart, and Winston.
Hymes, D. H. (1990). Epilogue to 'The things we do with words'. In D. Carbaugh (Ed.), *Cultural communication and intercultural contact* (pp. 419–429). Hillsdale, NJ: Lawrence Erlbaum.
Johnson, D. C. (2004). Language policy discourse and bilingual language planning. *Working Papers in Educational Linguistics, 19*(2), 73–97.
Knobel, M. (1999). *Everyday literacies: students, discourse, and social practice*. New York, NY: Peter Lang.
Kress, G. (1997). *Before writing : Rethinking the paths to literacy.* London: Routledge.
Kress, G. & van Leeuwen, T. (1996). *Reading images: The grammar of visual design*. London: Routledge.

Lin, A. & Martin, P. (Eds.). (2005). *Decolonisation, globalisation: Language-in-education policy and practice.* Clevedon, UK: Multilingual Matters.

Martin-Jones, M. & Jones, K. (Eds.). (2000). *Multilingual literacies: Reading and writing different worlds*. Amsterdam: John Benjamins.

McCarty, T. L. (Ed.). (2005). *Language, literacy, and power in schooling*. Mahwah, NJ: Lawrence Erlbaum.

McLaughlin, D. (1992). *When literacy empowers: Navajo language in print*. Albuquerque, NM: University of New Mexico Press.

Page, R. (2000). The turn inward in qualitative research. In B. M. Brizuela, J. P. Stewart, et al. (Eds.), *Acts of inquiry in qualitative research* (pp. 3–16). Cambridge, MA: Harvard Educational Review.

Pérez, B., Flores, B., & Strecker, S. (2003). Biliteracy teacher education in the U.S. Southwest. In N. H. Hornberger (Ed.), *Continua of biliteracy: An ecological framework for educational policy, research, and practice in multilingual settings* (pp. 207–231). Clevedon, UK: Multilingual Matters.

Pennycook, A. (1994). Incommensurable discourses? *Applied Linguistics, 15*(2), 115–138.

Philips, S. U. (1983). *The invisible culture: Communication in classroom and community on the warm springs reservation*. New York, NY: Longman.

Pike, K. (1954). Emic and etic standpoints for the description of behavior. In K. Pike (Ed.), *Language in relation to a unified theory of the structure of human behavior* (pp. 37–72). The Hague: Mouton.

Rampton, B. (2000). Speech community. In J. Verschueren, J. Östman, J. Blommaert, & C. Bulcaen (Eds.), *Handbook of pragmatics 1998* (pp. 1–34). Amsterdam: John Benjamins.

Ricento, T. K. (Ed.). (2006). *An introduction to language policy: Theory and method*. New York, NY: Blackwell Publishing.

Ricento, T. K. & Hornberger, N. H. (1996). Unpeeling the onion: Language planning and policy and the ELT professional. *TESOL Quarterly, 30*(3), 401–428.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, *26,* 99–128.

Sipe, L. R. & Ghiso, M. P. (2004). Developing conceptual categories in classroom descriptive research: Some problems and possibilities. *Anthropology and Education Quarterly, 35*(4), 472–485.

Spener, D. (Ed.). (1994). *Adult biliteracy in the United States*. Washington, DC: Center for Applied Linguistics.

Stake, R. (2000). Case studies. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (pp. 435–454). Thousand Oaks, CA: Sage.

Stein, P. (2004). Re-sourcing resources: Pedagogy, history and loss in a Johannesburg classroom. In M. R. Hawkins (Ed.), *Language learning and teacher education: A sociocultural approach* (pp. 35–51). Clevedon, UK: Multilingual Matters.

Tollefson, J. W. (Ed.). (1995). *Power and inequality in language education*. New York, NY: CUP.

Tollefson, J. W. (Ed.). (2002). *Language policies in education: Critical issues*. Mahwah, NJ: Lawrence Erlbaum.

Tollefson, J. W. & Tsui, A. B. M. (Eds.). (2004). *Medium of instruction policies: Which agenda? Whose agenda?* Mahwah, NJ: Lawrence Erlbaum.

Wallace, A. F. C. (1961). *Culture and personality*. New York, NY: Random House.
Wortham, S. & Rymes, B. (Eds.). (2003). *Linguistic anthropology of education*. Westport, CT: Praeger.

Author index

Subject index

In the series *Language Learning & Language Teaching* the following titles have been published thus far or are scheduled for publication:

13 **NORRIS, John M. and Lourdes ORTEGA (eds.):** Synthesizing Research on Language Learning and Teaching. xiv, 341 pp. + index. *Expected August 2006*

12 **CHALHOUB-DEVILLE, Micheline, Carol A. CHAPELLE and Patricia A. DUFF (eds.):** Inference and Generalizability in Applied Linguistics. Multiple perspectives. 2006. vi, 248 pp.

11 **ELLIS, Rod (ed.):** Planning and Task Performance in a Second Language. 2005. viii, 313 pp.

10 **BOGAARDS, Paul and Batia LAUFER (eds.):** Vocabulary in a Second Language. Selection, acquisition, and testing. 2004. xiv, 234 pp.

9 **SCHMITT, Norbert (ed.):** Formulaic Sequences. Acquisition, processing and use. 2004. x, 304 pp.

8 **JORDAN, Geoff:** Theory Construction in Second Language Acquisition. 2004. xviii, 295 pp.

7 **CHAPELLE, Carol A.:** English Language Learning and Technology. Lectures on applied linguistics in the age of information and communication technology. 2003. xvi, 213 pp.

6 **GRANGER, Sylviane, Joseph HUNG and Stephanie PETCH-TYSON (eds.):** Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. 2002. x, 246 pp.

5 **GASS, Susan M., Kathleen BARDOVI-HARLIG, Sally Sieloff MAGNAN and Joel WALZ (eds.):** Pedagogical Norms for Second and Foreign Language Learning and Teaching. Studies in honour of Albert Valdman. 2002. vi, 305 pp.

4 **TRAPPES-LOMAX, Hugh and Gibson FERGUSON (eds.):** Language in Language Teacher Education. 2002. vi, 258 pp.

3 **PORTE, Graeme Keith:** Appraising Research in Second Language Learning. A practical approach to critical analysis of quantitative research. 2002. xx, 268 pp.

2 **ROBINSON, Peter (ed.):** Individual Differences and Instructed Language Learning. 2002. xii, 387 pp.

1 **CHUN, Dorothy M.:** Discourse Intonation in L2. From theory and research to practice. 2002. xviii, 285 pp. (incl. CD-rom).