VALIDATION IN LANGUAGE ASSESSMENT

Edited by

Antony John Kunnan

VALIDATION IN LANGUAGE ASSESSMENT

Selected Papers from the 17th Language Testing Research Colloquium, Long Beach

VALIDATION IN LANGUAGE ASSESSMENT

Selected Papers from the 17th Language Testing Research Colloquium, Long Beach

Edited by

Antony John Kunnan

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First Published by
Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, New Jersey 07430

Transferred to Digital Printing 2009 by Routledge
270 Madison Ave, New York NY 10016
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Cover design by Kathryn Houghtaling Lacey

Library of Congress Cataloging-in-Publication Data

Language Testing Research Colloquium (17th : 1995 : Long Beach, Calif.)
Validation in language assessment : selected papers from the 17th Language Testing Research Colloquium, Long Beach / edited by Antony John Kunnan.
p. cm.
Selected papers from the 17th annual colloquium held Mar. 24–27, 1995, in Long Beach, Calif.
Includes bibliographical references and indexes.
ISBN 0-8058-2752-8 (cloth : alk. paper). — ISBN 0-8058-2753-6 (pbk. : alk. paper).
1. Language and languages—Ability testing—Congresses. 2. Language and languages—Examinations—Congresses. I. Kunnan, Antony John. II. Title.
P53.4.L38 1995
418'.0076—dc21 98-25643
CIP

Publisher's Note
The publisher has gone to great lengths to ensure the quality of this reprint but points out that some imperfections in the original may be apparent.

To

Professor Jacob Tharu
of the Central Institute of English and Foreign Languages, Hyderabad,
who first introduced me to language test validation

Contents

PART II: TEST-TAKER CHARACTERISTICS AND FEEDBACK

PART III: PERSPECTIVE ON VALIDATION RESEARCH

Preface

Although validation of language (second and foreign language) assessment instruments is considered a necessary technical component of test design, development, maintenance, and research as well as a moral imperative for all stakeholders who include test developers, test-score users, test stockholders, and test-takers, only recently have language assessment researchers started using a wide variety of validation approaches and analytical and interpretive techniques.

This volume, which is made up of selected papers from the 17th Language Testing Research Colloquium, the premier annual international conference, contributes to this variety by presenting diverse approaches with an international perspective of validation in language assessment. The volume opens with an introduction to approaches to validation in language assessment in published research in the last 15 years. This is followed by 11 chapters in 3 sections: Part I presents four papers that focus on validation through the stages of test development and test-taking process. Part II presents six papers that focus on validation by examining data from test-taker characteristics and test-taker feedback. Part III presents an analytical assessment of the presentations at 15 Language Testing Research Colloquiums. In all, the 12 chapters provide excellent examples of the different approaches language assessment researchers have taken to validation. In addition, the international perspective offered coupled with an annotated suggested readings list after each chapter should interest a wide variety of individuals interested in validation of language assessment instruments:

graduate and doctoral students of language assessment and evaluation, educational researchers, and government administrators and policymakers.

This volume has benefited from many who contributed careful thoughts, exemplary diligence, and above all, saintly patience. Obviously, I would like to thank all the contributors not only for taking the time to write their papers, but also for taking the time again to revise them in the light of my views as to the nature the volume should take. Less obviously, I am grateful to Lyle Bachman for his invaluable encouragement and advice toward this project, to the three reviewers for providing insightful comments on all the papers, and, to Naomi Silverman of Lawrence Erlbaum Associates for her quiet persuasion and commendable attention to quality, without whose interest this volume would not have been published. To all who were present at LTRC '95 in Long Beach, I want to thank you for waiting patiently for this volume. I hope it is just in time to bring back many memories, including a quiet night of stars aboard the Queen Mary!

Antony John Kunnan

Foreword

Lyle F. Bachman
University of California, Los Angeles

The 17th annual Language Testing Research Colloquium (LTRC) was held March 24–27, 1995, in Long Beach, California, with the theme "Validity and Equity Issues in Language Testing." The plenary address entitled "Validity and Equity Issues in Educational Assessment" was given by Eva Baker, Director of the University of California, Los Angeles' Center for Studies in Evaluation, and Co-Director of the National Center for Research, Evaluation, Standards, and Student Testing. This was followed by a panel discussion of validity and equity issues in assessment by distinguished scholar/researchers from the fields of language testing, educational measurement, and educational policy. This opening session set the theme and tone for the presentation of many outstanding papers presented, 11 of which are included in this volume.

The theme of validity is no stranger to the LTRC, as can be seen in the paper by Liz Hamp-Lyons and Brian Lynch in this volume. Indeed, validation has been a major thread running through virtually all the LTRCs since the first in 1979, whose stated theme was "The Construct Validation of Tests of Communicative Competence." That first LTRC grew out of the ferment brought about by the confluence of two differing views of language ability and their implications for language testing: John Oller's unitary trait hypothesis and Mike Canale and Merrill Swain's multicomponential view of communicative competence. As Bachman and Palmer (1988) pointed out in their introduction to the special issue of *Language Testing* devoted to papers from the 10th LTRC, one of "the focal points that emerged from the first

LTRC was an interest in a broader view of language proficiency as communicative competence" (p. 126). Also emerging from that first LTRC was "a determination to embark on a program of empirical research into the then relatively unknown realm of construct validation" (Bachman & Palmer, 1988, p. 126). Thus, even though the papers presented at that first LTRC were relatively unsophisticated statistically, by today's standards, they raised many of the validity issues, both conceptual and methodological, surrounding the nature of language ability and its measurement that are still with us today.

The continuing LTRC interest in and concern with investigating validity issues in language testing is reflected in the fact that validation has been the theme of six subsequent colloquia, the 2nd (1980), the 3rd (1981), the 4th (1982), the 10th (1988), the 14th (1992), and the 17th (1995). LTRC's interest in validity issues has had, I believe, a substantial influence on the field itself. Papers presented at the LTRC regularly appear in *Language Testing* and other professional journals in applied linguistics. In addition, nine volumes, including this one, of selected LTRC papers have been published over the years, adding an invaluable resource to the research literature now available to language testers. Individuals interested in language testing research may also log on to the International Language Testing Association (ILTA) home page at http://www.surrey.ac.uk/ELI/ilta/ilta.html to access all past LTRC program books, which include paper abstracts and many of the papers themselves.

Our understanding of validity and the process of validation has deepened since the first LTRC in 1979, as is discussed in Antony Kunnan's introduction to this volume, and demonstrated, I believe, by the bibliographic entries in the Appendix to his introduction. This deepened understanding reflects both the expanded view of validity in educational measurement and an awareness that language testing presents validity conundrums of its own. At the same time, our technical and methodological toolbox has expanded, so that it is now commonplace to see LTRC papers and journal articles in language testing that employ computer-based and multimedia approaches to test design and administration, as well as structural equation modeling, many-facet Rasch and generalizability theory in the analysis of test results and the validation process. The role of the LTRC in expanding our awareness of the immense scope of validation research, will, I believe, continue to be vital. An increasing number of researchers from other areas of applied linguistics and from educational measurement are attending the LTRC on a regular basis, providing the opportunity for greater links between language testers and researchers in these fields. I am also confident that published volumes of LTRC papers, including this one, as well as those from subsequent LTRCs, will continue to define the cutting edge of validation research in language testing.

REFERENCE

Bachman, L. F., & Palmer, A. S. (1988). 10th annual LTRC programme: Programme introduction. *Language Testing, 5*, 125–127.

CHAPTER

1

Approaches to Validation in Language Assessment

Antony John Kunnan
California State University, Los Angeles

Since the 1960s, the central location of intense language assessment (and testing) research has been validation. In the 1960s and 1970s, language assessment developers and researchers, like fellow educational and psychological testing and measurement researchers (Angoff, 1988), initially followed the rather narrow 1954 and 1966 Standards for Educational and Psychological Tests and Manuals (American Psychological Association [APA], 1954, 1966) and the Cronbach and Meehl (1955) proposal, devoting their attention to the five traditional types of validity: face-content, criterion-related, predictive, concurrent, and construct. Evidence of this segmented approach to validation can be seen in the numerous language assessment research studies reported in journals and textbook chapters of this period and even later (Alderson, Clapham, & Wall, 1995; Hughes, 1989; Lado, 1961).

In 1985, the revised Standards (APA, 1985) was published, and its greatly expanded view of validity included testing standards for different purposes, contexts, and groups. It also asserted that validity is a unitary concept, referring to the "appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p. 8). About this time, Messick (1980, 1989) presented his fully articulated thoughts regarding a unified validity framework. He (1989) asserted that a unified validity framework could be constructed

> by distinguishing two interconnected facets of the unitary validity concept. One facet is the source of justification of the testing, being based on appraisal

TABLE 1.1
Progressive Matrix View of Validity

	Test Interpretation	*Test Use*
Evidential basis	Construct validity	Construct validity + Relevance/utility
Consequential basis	Construct validity + Value implications	Construct validity + Relevance/utility + Value implications + Social consequences

From Messick (1989).

> of either evidence or consequence. The other facet is the function or the outcome of the testing, being either interpretation or use. If the facet for source of justification (that is either an evidential basis or a consequential basis) is crossed with the function or outcome of the testing (that is, either test interpretation or test use), we obtain a four-fold classification. (p. 20)

When Messick's framework is read as a progressive matrix with the different facets contributing to this unified validity concept, the overall influence of construct validity and the critical importance of each facet become clearer. This progressive matrix view of Messick's fourfold classification of facets of validity is presented in Table 1.1.

This was the first time concepts such as value implications and social consequences were introduced within the framework of assessment validation, offering the scope and possibility of including constructs of social and cultural difference and social consequences in validation research. Although this unified framework has been widely accepted by educational and language assessment and testing researchers and today is the cornerstone for most validation research, not all aspects of this framework have received equal research emphasis, and these gaps will be noticeable in subsequent sections of this introduction.

IMPLICATIONS OF MESSICK'S FRAMEWORK FOR LANGUAGE ASSESSMENT

Implications of Messick's framework specifically for language assessment and test developers and researchers was first outlined by Bachman (1990). Under the category of evidential basis for validity, he lists five different types of empirical evidence that can be collected in support of construct validity. Bachman stated that the most powerful types of evidence are correlational evidence regarding item scores and test scores (and by default, language proficiency and test dimensionality) as well as experimental evidence re-

garding the effects of experimental treatment. Other types of evidence he listed include analyses of test-taking processes, studies of group differences among test takers, and studies of changes over time.

Under the category of consequential basis for validity, Bachman lists four areas to be considered in the interpretation and use of test scores: (a) construct validity, or the evidence that supports the particular interpretation; (b) multiple perspectives on value systems from test takers, test developers, and test users; (c) practical usefulness of tests; and (d) misuse of tests, the ethics of test use, and the social consequences of test invalidity to society. Bachman also argued that it is necessary to consider alternatives to testing as an area of examination of social consequences of testing.

Organizing these themes as listed by Bachman in Table 1.2 would, on the one hand, clearly show the different lines of inquiry that are possible and necessary and, on the other, show how the different lines of inquiry fit together into the unified concept of test validation.

VALIDATION STUDIES IN LANGUAGE ASSESSMENT

Translating the language assessment and testing research themes presented in Table 1.2 into key research projects that have engaged language assessment researchers will shed light on the areas where the focus has been and perhaps adequate understanding of issues does exist, where there are gaps, and where more attention is needed. A survey of assessment validation research in the post-1980 period was conducted for this purpose.

Table 1.3 presents the names and years of the researchers and research studies organized in the Messick framework; citations are presented in the

TABLE 1.2
Language Assessment Research Themes in Messick's Framework

	Test Interpretation	*Test Use*
Evidential basis	1. Proficiency components 2. Test dimensionality 3. Test-validation process 4. Test development: New test methods, rating scales, conditions, etc.	1. Test-taking processes 2. Test-taking strategies 3. Test-taker characteristics: Academic background, native language and culture, field in/dependence; DIF studies: native language and culture, gender, ethnicity, age, etc.
Consequential basis	1. Value system differences: Test-taker and specialists' feedback	1. Social consequences and washback 2. Ethics, standards and equity 3. Alternatives

TABLE 1.3
Key Language Assessment Studies in Messick's Framework (1980-1996)

Test Interpretation	*Test Use*
Evidential Basis	
1. Proficiency components Oller & Hinofotis, 1980 Flahive, 1980 Scholz et al., 1980 Bachman & Palmer, 1981, 1982 Carroll, 1983 Oller, 1983 Hinofotis, 1983 Upshur & Homburg, 1983 Vollmer & Sang, 1983 Sang et al., 1986 Hale, 1989 Turner, 1989	**1. Test-taking processes** Alderson, 1990a, 1990b Buck, 1991 Perkins, 1992 Ross, 1992 Lumley, 1993 Rost, 1993 Freedle & Kostin, 1993 Hale & Courtney, 1994 Stansfield & Kenyon, 1996
2. Test dimensionality Henning et al., 1985 de Jong & Glas, 1987 Davidson, 1988 Boldt, 1989, 1992 Henning, 1992 McNamara, 1991 Choi & Bachman, 1992 Blais & Laurier, 1995	**2. Test-taker strategies** Anderson et al., 1991 Wijh, 1996
3. Test-validation process Davies, 1984 Clark, 1988 Bachman et al., 1988, 1995 McNamara, 1990 Shohamy & Inbar, 1991 Kunnan, 1992 Shohamy, 1994 Scott et al., 1996 Cumming & Mellow, 1996	**3. Test-taker characteristics:** **Academic background** Alderson & Urquhart, 1985 Chihara et al., 1989 Hale, 1988 Clapham, 1993, 1996 Jensen & Hansen, 1995 **Native language/culture, gender, ethnicity, age** Swinton & Powers, 1980 Alderman & Holland, 1981 Chen & Henning, 1985 Zeidner, 1986, 1987 Oltman et al., 1988 Duran, 1988 Angoff, 1989 Kunnan, 1990, 1994 Ryan & Bachman, 1992 **Field in/dependence** Stansfield & Hansen (Ross), 1983 Hansen (Ross) & Stansfield, 1984 Chapelle, 1988
4. Test development: Cloze, c-test, translation, summary, vocabulary Alderson, 1989 Bachman, 1982 Brown, 1987, 1993 Chapelle & Abraham, 1990 Jonz, 1991 Klein-Braley, 1985 Buck, 1992 Huhta & Rendell, 1996 Read, 1993 **Scales** Davidson & Henning, 1985 Chalhoub-Deville, 1995 Milanovic et al., 1996 Tyndall & Kenyon, 1996	

(Continued)

TABLE 1.3
(Continued)

Contexts Hamp-Lyons, 1991 Spaan, 1993 Cohen, 1993	
Consequential Basis	
1. **Value system differences: Test-taker feedback** Cohen, 1984 Zeidner & Bensoussan, 1988 Bradshaw, 1990 Brown, 1993 Peirce & Stein, 1995 **Specialists' feedback** Elder, 1993	1. **Social consequences: Impact & Washback effect** Wall & Alderson, 1993 Messick, 1996 2. **Ethics, Standards, Equity** Spolsky, 1981 Stansfield, 1993 Tharu, 1993 3. **Alternatives** Oscarsson, 1989 Heilenman, 1990 Hamayan, 1995

Appendix at the end of this chapter. Here is a brief description and comment on this list of key studies. In the Test Interpretation section under Evidential Basis, the focus is on four research areas: language proficiency components, test dimensionality, test validation process, and test development.

Studies on *language proficiency* primarily investigated whether language proficiency was multicomponential, or unidimensional as was claimed by Oller and his colleagues (see Oller & Perkins, 1980). Researchers vigorously pursued this question using several methodologies such as correlational analysis, exploratory and confirmatory factor analysis, and multitrait–multimethod design with test performance data from different tests and contexts. They found satisfactory and convincing evidence from the analyses that language proficiency was multicomponential, not unidimensional. This led to the conviction, for most researchers at least, that language proficiency is multicomponential.

This line of research shifted in the late 1980s to capture a second perspective on the same issue: *test dimensionality*. Analyses were predominantly conducted with different applications of Item Response Theory (IRT), such as the Rasch Model, the two- and three-parameter models, and the Bejar procedure. Despite the intense activity in these areas, the main questions regarding the specific components of the multicomponentiality of proficiency and dimensionality of language tests have not been unambiguously answered.

The third line of research, pursued with much persistence, has been finding evidence for *test validation* by analyzing test performance data. These studies have included analyses of tasks and abilities, oral and written language samples, reading texts and question types, and test scores. Different methodologies have been used including content analysis, factor and cluster analysis and generalizability theory, and IRT.

The fourth line of research, *test development* research, has seen much activity, specifically in examining new tests, new rating scales and experimental test conditions. Findings from these studies have been valuable to both test developers and researchers.

The key studies in the Test Use category in the Evidential Basis section fall into three categories: test-taking process, test-taking strategies, and test-taker characteristics. The studies in the *test-taking process* category have examined various test-taking processes in skill areas such as reading and listening comprehension, oral proficiency, and scaling of speaking tasks. The *test-taking strategies* studies focus on strategies used by those taking tests. As is obvious from the few studies in this category, much more needs to be done so that there can be a better understanding of test-taking strategies deployed by test takers in different test contexts. The studies under the category of *test-taker characteristics* have focused on test takers' academic background, native language and culture, gender, ethnicity, field in/dependence and differentially item functioning. This area of investigation has generated awareness among test developers and researchers that test takers from certain social, cultural, academic, native language and culture, gender, ethnicity, age, and learning style groups might be affected by a test or its items in ways that are not relevant to the abilities being tested. Moreover, these studies have also been in the forefront of asking whether tests or items and score use are fair to all test takers.

The list of studies in the Consequential Basis section compared to the list in the Evidential Basis section is smaller and more recent; it is here that the yawning gaps lie. Under the Test Interpretation category, the small number of studies have focused primarily on obtaining feedback from test takers regarding tests they have taken. This type of research has recently also included feedback from college or university subject matter specialists on tests. Both groups have been able to provide opinions regarding test content, test format/method, test process, and test appropriacy, all of which up till now were assumed to be known by test developer and researchers.

Under the Test Use category, three areas of interest have developed: social consequences, mainly washback effect; ethics, equity, and standards; and alternatives to tests. Although the topic of *washback effect* has been discussed in language testing for many years, systematic attempts to understand the phenomenon were made only recently. The few studies on *ethics and standards* on the one hand have focused on the need for responsibility

and accountability in language testing, and on the other have targeted technically appropriate procedures or standards for test developers and agencies in test development, interpretation, and score use. Finally, under the general area of research called *alternatives*, a few researchers have typically focused on self-assessment as an alternative way of doing what tests typically do.

Presenting these studies in Messick's framework offers an examination of the different research themes in assessment validation that have been investigated over the past 16 years. This presentation reveals an imbalance in the attention researchers have given these facets of Messick's framework. Test Interpretation in the Evidential Basis section has received the most attention and is clearly the conventional approach in examining test validation. Test Use in the Evidential Basis section has received more recent attention, and Test Interpretation and Test Use in the Consequential Basis section is just beginning to receive attention. In general, this imbalance has to be corrected. Furthermore, the approach used in these last three areas is in contrast to the conventional approach and can perhaps be termed a postmodern approach in examining test validation. Moreover, if the language assessment and testing community is committed to understanding its place in postmodern societies, then it is not just the unbalanced approach to assessment validation that needs to be reexamined, but a proactive research agenda that focuses on Test Interpretation and Test Use under Consequential Basis has to be formulated (see Kunnan, 1997 for an argument connecting fairness with validation).

THEMES FROM INDIVIDUAL CHAPTERS

The chapters in this volume in many ways further our understanding of assessment validation approaches that belong to both the conventional and the postmodern approach, although they focus more on the latter approach. Following Cumming's (1996) model, Table 1.4 presents the chapters in this volume under Messick's unified framework.

The 11 chapters that follow are presented in three sections: test development and test-taking process (4 chapters), test-taker characteristics and feedback (6 chapters), and general validation (1 chapter).

Dorry Kenyon's chapter leads the discussions in section I, which illustrates the conventional approach to assessment validation research. Kenyon investigates foreign language students' perceived difficulty in performing various speaking tasks in a manner consistent with the hierarchical characterizations of these tasks in the Speaking Proficiency Guidelines of the American Council on the Teaching of Foreign Languages. Using language test performance data from high school and college students in French,

TABLE 1.4
Themes and Chapters in This Volume, Following Messick's Framework

Test Interpretion	*Test Use*
Evidential Basis	
Test development Read, chapter 3 Fortus et al., chapter 4 Wiggelsworth, chapter 5	**Test-taking process** Kenyon, chapter 2 **Test-taker characteristics** Purpura, chapter 6 Clapham, chapter 7 Ginther & Stevens, chapter 8 Brown & Iwashita, chapter 9 Hill, chapter 10
Consequential Basis	
Test-taker feedback Norton & Stein, chapter 11	
Perspective on Validation Research	
Kunnan, chapter 1 Hamp-Lyons & Lynch, chapter 12	

German, and Spanish, Kenyon employed the many-facet Rasch model for his analysis.

Three chapters on test development follow, each one focusing on a single concern: Read on a new test format; Fortus, Coriat, and Fund on item difficulty; and Wiggelsworth on the special test condition of planning. John Read's chapter focuses on validating the word associates' format as a measure for depth of vocabulary knowledge with test performance data from New Zealand. This word associates' format essentially bridges the gap between measures of breadth and depth of vocabulary knowledge. He also used concurrent measures such as a matching test and an interview to provide evidence for concurrent validity of the new format.

The chapter by Ruth Fortus, Rikki Coriat, and Susan Fund examines the difficulty levels of items in the reading section of an English test in Israel so that test developers can design item pools in accordance with specific needs, such as items for low and high abilities. Isolating the factors that affect item difficulty, they argue, will increase test developers' understanding of the construct validity of the test.

Gillian Wiggelsworth's chapter focuses on an important, though neglected, aspect of a test: planning time. Her chapter discusses the effect of planning time on second language oral test discourse in a semidirect oral interaction test in Australia. Using discourse analytic techniques, she examines the nature and significance of differences in elicited discourse across the two conditions of the test in terms of fluency, accuracy, and complexity in the second language.

Section II of this volume focuses on test-taker characteristics and feedback, which illustrates a less conventional approach and arguably a postmodern approach to assessment validation research. James Purpura's chapter presents the development and construct validation of a cognitive processes questionnaire instrument designed to investigate selected cognitive background characteristics of test takers in the United States. The processes considered in this instrument validation study were selecting, comprehending, storing/memory, and using or retrieval. Purpura uses exploratory factor analysis procedures in the development of a taxonomy of cognitive strategies.

Caroline Clapham's chapter examines the effect of language proficiency and background knowledge on students' reading comprehension in the United Kingdom. The aim of Clapham's study is to consider the effect of background knowledge on reading comprehension, and to examine whether an English for specific approach to testing was appropriate. Subject matter and topic familiarity, language proficiency, and level of specificity of topics are variables the author examines with regression and analysis of variance procedures.

The chapter by April Ginther and Joseph Stevens investigates the internal construct validity of an advanced placement Spanish language examination. The authors compare the factor structure of the test of Latin Spanish-speaking test takers with those of Mexican Spanish-speaking, Mexican Spanish-English bilingual, White English-speaking, and Black English-speaking groups in the United States. Implications of differences in factor structure, loadings, and variances are valuable for test developers as well as for test validity research.

Annie Brown and Noriko Iwashita's chapter examines the role of native language background in the validation of a computer-adaptive test. The authors use test performance data from beginning to intermediate students of Japanese and from native speakers of English, Chinese, and Korean on a 225-item multiple-choice test of grammar to identify item difficulties. The difficulty of the items was discovered to be different for the three native language, and the ramifications of this finding for the validation of a computer-adaptive test are discussed.

Kathryn Hill's chapter investigates the effect of test-taker characteristics on reactions to an oral English proficiency test. Using feedback from different groups of test takers such as Asians and Europeans, male and female,

and students and professionals in Australia, Hill examines questionnaire responses to an oral proficiency test using FACETS, the multifaceted Rasch analysis program.

The last chapter presented in terms of Messick's consequential basis is Bonny Norton and Pippa Stein's chapter, which addresses issues of textual meaning, testing, and pedagogy on the basis of their experience piloting a college entrance reading test in English for Black students in South Africa. The authors discuss how the students' interpretations of the reading comprehension text differed from that of the test developers. They also raise probing questions at the heart of testing, equity, and pedagogy.

As a fitting conclusion to the volume, Liz Hamp-Lyons and Brian Lynch examine research practices of the second- and foreign-language testing community as seen through the Language Testing Research Colloquium series in the last 15 years. The authors focus their analysis on the ways in which test validity and reliability have been addressed both implicitly and explicitly in language testing research. Furthermore, their inquiry explores whether traditional psychometric approaches or newer alternative perspectives and modes of inquiry as suggested in recent measurement literature are used by language testing researchers.

In summary, although these chapters have brought to light the critical themes of language assessment validation through conventional and perhaps postmodern approaches, there are many areas of investigation worthy of attention that are not represented here. These include less popular topics such as standards, equity, and alternatives to testing, as well as traditional topic areas such as proficiency components and test dimensionality.

Furthermore, the Messick test validation framework itself might be characterized as a rather conventional approach to assessment validation. Indeed, it presents just one view of how assessment validation can be conceptualized, researched, and reported. Compelling alternative perspectives deserve serious attention and wider recognition, such as the hermeneutic approach to validation proposed by Moss (1994) and a much more radical and political approach proposed by Cherryholmes (1988) who argues that critical research and history must be represented in validation attempts, as in his opinion, "construct validity decisions are ethico-political and aesthetic as well as social scientific" (p. 127). These alternative approaches will most certainly add to our understanding of assessment validation, even though they could signal a radical departure from the Cronbach and Meehl validation approach of 1955 and the Messick approach of the 1980s that this introduction has traced. Future volumes hopefully will track the conventional, the postmodern, and the radical approaches deployed by language assessment researchers, raising both public and professional awareness regarding assessment validation and resulting in responsible test use in all contexts and for all concerned.

APPENDIX: KEY LANGUAGE ASSESSMENT STUDIES

Test Interpretation: Evidential Basis

Proficiency Components

Oller, J., & Hinofotis, F. (1980). Two mutually exclusive hypotheses about second language ability: Indivisible or partially divisible competence. In J. Oller (Ed.), *Research in language testing* (pp. 13–23). Rowley, MA: Newbury House.

Flahive, D. (1980). Separating the g factor from reading comprehension. In J. Oller (Ed.), *Research in language testing* (pp. 34–46). Rowley, MA: Newbury House.

Scholz, G., Hendricks, D., Spurling, R., Johnson, M., & Vandenburg, L. (1980). Is language ability divisible or unitary? A factor analysis of 22 English language proficiency tests. In J. Oller (Ed.), *Research in language testing* (pp. 24–33). Rowley, MA: Newbury House.

Bachman, L. F., & Palmer, A. (1981). The construct validation of the FSI oral interview. *Language Learning, 31*, 67–86.

Bachman, L. F., & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449–465.

Carroll, J. (1983). Psychometric theory and language testing. In J. Oller (Ed.), *Issues in language testing research* (pp. 80–107). Rowley, MA: Newbury House.

Oller, J. (1983). Evidence for a general language proficiency factor: An expectancy grammar. In J. Oller (Ed.), *Issues in language testing research* (pp. 3–10). Rowley, MA: Newbury House.

Hinofotis, F. (1983). The structure of oral communication in an educational environment: A comparison of factor analytic rotational procedures. In J. Oller (Ed.), *Issues in language testing research* (pp. 170–187). Rowley, MA: Newbury House.

Upshur, J., & Homburg, T. (1983). Some relations among language tests at successive ability levels. In J. Oller (Ed.), *Issues in language testing research* (pp. 188–202). Rowley, MA: Newbury House.

Vollmer, H., & Sang, F. (1983). Competing hypotheses about second language ability: A plea for caution. In J. Oller (Ed.), *Issues in language testing research* (pp. 29–79). Rowley, MA: Newbury House.

Sang, F., Schmitz, B., Vollmer, H., Baumert, J., & Roeder, P. (1986). Models of second language competence: A structural equation modeling approach. *Language Testing, 3*, 54–79.

Hale, G., Rock, D., & Jirele, T. (1989). Confirmatory factor analysis of the TOEFL. TOEFL Research Reports 32. Princeton, NJ: Educational Testing Service.

Turner, C. E. (1989). The underlying factor structure of L2 cloze test performance in francophone, university-level students: Causal modeling as an approach to construct validation. *Language Testing, 6*, 172–197.

Test Dimensionality

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing, 2*, 141–154.

de Jong, J., & Glas, C. (1987). Validation of listening comprehension tests using item response theory. *Language Testing, 4*, 170–194.

Davidson, F. (1988). *An exploratory modeling survey of the trait structure of some existing language testing data sets.* Doctoral dissertation, University of California, Los Angeles.

Boldt, R. (1989). Latent structure analysis of the TOEFL. *Language Testing, 6*, 125–142.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing, 8*, 45–65.

Boldt, R. (1992). Crossvalidation of item response curve models using TOEFL data. *Language Testing, 9*, 79–95.

Henning, G. (1992). Dimensionality and construct validity of tests. *Language Testing, 9*, 1–11.

Choi, I.-C., & Bachman, L. (1992). An investigation into the adequacy of three IRT models of data from two EFL reading tests. *Language Testing, 9*, 51–78.

Blais, J.-G., & Laurier, M. (1995). The dimensionality of a placement test from several analytical perspectives. *Language Testing, 12*, 72–98.

Test Validation Process

Davies, A. (1984). Validating three tests of English language proficiency. *Language Testing, 1*, 50–69.

Clark, J. (1988). Validation of a tape-mediated ACTFL/ILR scale-based test of Chinese speaking proficiency. *Language Testing, 5*, 187–205.

Bachman, L., Kunnan, A., Vanniarajan, S., & Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing, 5*, 128–159.

McNamara, T. F. (1990). IRT and the validation of an ESP test for health professionals. *Language Testing, 7*, 52–75.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension types: The effect of text type and question type. *Language Testing, 8*, 23–40.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analysis. *Language Testing, 9*, 30–49.

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

Bachman, L., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of EFL: The Cambridge-TOEFL comparability study*. Cambridge, UK: Cambridge University Press.

Scott, M. L., Stansfield, C., & Kenyon, D. (1996). Examining validity in a performance test: The listening summary translation exam (LSTE) -Spanish version. *Language Testing, 13*, 83–101.

Cumming, A., & Mellow, D. (1996). An investigation into the validity of written indicators of second language proficiency. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 72–93). Clevedon, UK: Multilingual Matters.

Test Development

Cloze, C-Test

Alderson, J. C. (1980). Native and non-native speaker performance on cloze tests. *Language Learning, 30*, 59–76.

Bachman, L. (1982). The trait structure of cloze test scores. *TESOL Quarterly, 16*, 61–70.

Brown, J. D. (1987). Tailored cloze: Improved with classical item analysis techniques. *Language Testing, 5*, 19–31.

Brown, J. D. (1993). What are the characteristics of *natural* cloze tests? *Language Testing, 10*, 93–116.

Chapelle, C., & Abraham, R. (1990). Cloze method: What difference does it make? *Language Testing, 7*, 121–146.

Jonz, J. (1991). Cloze item types and second language comprehension. *Language Testing, 8*, 1–22.

Klein-Braley, C. (1985). A cloze-up on the C-test: A study in the construct validation of authentic tests. *Language Testing, 2*, 76–118.

Translation, Summary, Vocabulary

Buck, G. (1992). Translation as a language testing procedure: Does it work? *Language Testing, 9*, 123–148.

Huhta, A., & Rendell, E. (1996). Multiple-choice summary: A measure of text comprehension. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 94–110). Clevedon, UK: Multilingual Matters.

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

Rating Scale Development

Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. *Language Testing, 2*, 164–191.

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

Milanovic, M., Saville, N., Pollitt, A., & Cook, A. (1996). Developing rating scales for CASE: Theoretical concerns and analyses. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 15–38). Clevedon, UK: Multilingual Matters.

Tyndall, B., & Kenyon, D. (1996). Validation of a new holistic rating scale using Rasch multifaceted analysis. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 39–57). Clevedon, UK: Multilingual Matters.

Prompts, Instructions

Hamp-Lyons, L., & Prochow, S. (1991). Prompt difficulty, task type, and performance. In S. Anivan (Ed.), *Current developments in language testing* (pp. 58–76). Singapore: SEAMEO/RELC.

Spaan, M. (1993). The effect of prompt in essay examinations. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 98–122). Alexandria, VA: TESOL.

Cohen, A. (1993). The role of instructions in testing summarizing ability. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 132–160). Alexandria, VA: TESOL.

Test Use: Evidential Basis

Test-Taking Processes

Alderson, J. C. (1990a). Testing reading comprehension skills. Part One. *Journal of Reading in a Foreign Language, 6*, 425–438.

Alderson, J. C. (1990b). Testing reading comprehension skills. Part One. *Journal of Reading in a Foreign Language, 7*, 465–503.

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing, 8*, 67–94.

Perkins, K. (1992). The effect of passage topical structure types on ESL reading comprehension difficulty. *Language Testing, 9*, 163–172.

Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing, 9*, 173–186.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing, 10*, 211–234.

Rost, D. (1993). Assessing the different components of reading comprehension: Fact or fiction? *Language Testing, 10*, 79–92.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing, 10*, 133–170.

Hale, G., & Courtney, R. (1994). The effects of note-taking on listening comprehension in the TOEFL. *Language Testing, 11*, 29–48.

Stansfield, C., & Kenyon, D. (1996). Comparing the scaling of speaking tasks by language teachers and by the ACTFL guidelines. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 124–153). Clevedon, UK: Multilingual Matters.

Test-Taking Strategies

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*, 41–66.

Wijh, I. (1996). A communicative test in analysis: Strategies in reading authentic texts. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 154–170). Clevedon, UK: Multilingual Matters.

Test-Taker Characteristics

Academic Background

Alderson, J. C., & Urquhart, A. (1985). The effect of students' academic discipline on EFL/ESL reading test takers. *Language Testing, 2*, 192–204.

Chihara, T., Sakurai, T., & Oller, J. (1989). Background and culture as factors in EFL reading comprehension. *Language Testing, 6*, 143–163.

Hale, G. (1988). Student major field and text content: Interactive effects on reading comprehension in the TOEFL. *Language Testing, 5*, 49–61.

Clapham, C. (1993). Is ESP testing justified? In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 257–271). Alexandria, VA: TESOL.

Jensen, C., & Hansen, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing, 12*, 99–119.

Clapham, C. (1996). What makes an ESP reading test appropriate for its candidates? In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 171–193). Clevedon, UK: Multilingual Matters.

Native Language, Culture, Gender, Etc.

Swinton, S. S., & Powers, D. E. (1980). *Factor analysis of the TOEFL for several language groups*. TOEFL Research Report 6. Princeton, NJ: Educational Testing Service.

Alderman, D., & Holland, P. (1981). *Item performance across native language groups on the TOEFL*. TOEFL Research Report 9. Princeton, NJ: Educational Testing Service.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*, 155–163.

Zeidner, M. (1986). Are English language aptitude tests biased towards culturally different minorities? Some Israeli findings. *Language Testing, 3*, 80–98.

Zeidner, M. (1987). A comparison of ethic, sex, and age bias in the predictive validity of English language aptitude tests: Some Israeli data. *Language Testing, 4*, 55–71.

Oltman, P., Stricker, J., & Barrows, T. (1988). *Native language, English proficiency, and the structure of the TOEFL.* TOEFL Research Report 27. Princeton, NJ: Educational Testing Service.

Duran, R. (1988/1993). Testing of linguistic minorities. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 573–587). London: Macmillan.

Angoff, W. (1989). *Context bias in TOEFL.* TOEFL Research Report 29. Princeton, NJ: Educational Testing Service.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly, 24*, 740–746.

Ryan, K., & Bachman, L. (1992). DIF on two tests of EFL proficiency. *Language Testing, 9*, 12–29.

Kunnan, A. J. (1994). Modelling relationships among some test-taker characteristics and performance on EFL tests: An approach to construct validation. *Language Testing, 11*, 225–252.

Field Dependence and Independence

Stansfield, C., & Hansen, J. (1983). Field dependence–independence as a variable in second language cloze test performance. *TESOL Quarterly, 17*, 29–38.

Hansen, J., & Stansfield, C. (1984). Field dependence–independence and language testing: Evidence from six Pacific-Island cultures. *TESOL Quarterly, 18*, 311–324.

Chapelle, C. (1988). Field independence: A source of language variation? *Language Testing, 7*, 121–146.

Test Interpretation: Consequential Basis

Value System Differences: Test Takers and Specialists

Cohen, A. (1984). On taking language tests: What the students report. *Language Testing, 1*, 70–81.

Zeidner, M., & Bensoussan, M. (1988). College students' attitudes towards written versus oral test of English as a foreign language. *Language Testing, 5*, 100–124.

Bradshaw, J. (1990). Test takers' reactions to a placement test. *Language Testing 7*, 13–30.

Brown, A. (1993). The role of test taker feedback in the test development process: Test takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*, 277–304.

Peirce, B., & Stein, P. (1995). Why the "Monkeys Passage" bombed: Tests, genre and teaching? *Harvard Educational Review, 65*, 50–65.

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10*, 235–254.

Test Use: Consequential Basis

Social Consequences and Washback

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*, 41–69.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256.

Ethics, Standards, and Equity

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-am-Main: Peter Lang.

Stansfield, C. (1993). Ethics, standards, and professionalism in language testing. *Issues in Applied Linguistics, 4*, 189–206.

Tharu, J. (1993). Tests of English proficiency: The problem of standards. *Journal of English as a Foreign Language, 3/4*, 59–78.

Alternatives

Oscarsson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing, 6*, 1–13.

Heilenmann, L. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing, 7*, 174–201.

Hamayan, E. (1995). Approaches to alternative assessment. *Annual Review of Applied Linguistics 15*, 212–226.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2), 2.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: Author.

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 19–32). Hillside, NJ: Lawrence Erlbaum Associates.

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

Cherryholmes, C. (1988). *Power and criticism: Poststructural investigations in education*. New York: Teachers College, Columbia University.

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

Cumming, A. (1996). Introduction: The concept of validation in language testing. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 1–14). Clevedon, UK: Multilingual Matters.

Hughes, A. (1989). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.

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

Lado, R. (1961). *Language testing*. London: McGraw-Hill.

Messick, S. (1980). Test validity and ethics of assessment. *American Psychologist, 35*, 1012–1027.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: Macmillan.

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

Oller, J. W., & Perkins, K. (1980). *Research in language testing*. Rowley, MA: Newbury House.