OXFORD APPLIED LINGUISTICS

Fundamental Considerations in Language Testing

Lyle F.Bachman

Oxford University Press

Fundamental Considerations in Language Testing

Lyle F. Bachman

Oxford University Press

Oxford University Press
Walton Street, Oxford OX2 6DP

Oxford New York
Athens Auckland Bangkok Bombay
Calcutta Cape Town Dar es Salaam Delhi
Florence Hong Kong Istanbul Karachi
Kuala Lumpur Madras Madrid Melbourne
Mexico City Nairobi Paris Singapore
Taipei Tokyo Toronto

and associated companies in
Berlin Ibadan

Oxford and *Oxford English* are trade marks of Oxford University Press

ISBN 0 19 437003 8

First published 1990
Third impression 1995

Typeset in 11 on 12 pt Sabon by Pentacor Ltd, High Wycombe, Bucks
Printed in Hong Kong

For my closest friends and loved ones: Nida, Tina, and Melissa

Acknowledgments

The author and publishers would like to thank the following for permission to reproduce the material below that falls within their copyright:

The American Council on Education for the extracts from R. L. Thorndike (ed.): ***Educational Measurement*** (Second edition) and R. L. Linn (ed.): ***Educational Measurement*** (Third edition)

The American Psychological Association for the extract from the paper by S. A. Messick in ***American Psychologist*** 30 .

Brooks/Cole Publishing Company for the table on page 30, adapted from M. J. Allen and W. M. Yen: ***Introduction to Measurement Theory***

The Center for Applied Linguistics for the extract from the pager by J. 0. Carroll in ***Testing the English Proficiency of Foreign Students***

Educational and Psychological Measurement for the extracts from the paper by C. I. Mosier in Volume 7

Jay Haley for the extract from ***Strategies of Psychotherapy***

The John Hopkins University Press for the figure on page 341, adapted from A. J. Nitko, 'Defining the criterion-referenced test' in R. A. Berk (ed.): A ***Guide to Criterion-Referenced Test Construction,*** and for the extract from the same paper

Language Learning, and the authors, for the tabie on page 196, from the paper by J. van Weeren and T. J. J. Theunissen in Volume 37

Newbury House Publishers for three figures (from J. A. Upshur, 'Context for language testing') in J. W. Oller and J. C. Richards: ***Focus*** on ***the Learner***

Pergamon Press PLC for the extract from K. Johnson (ed.): ***Communicative Syllabus Design and Methodology***

Psychometrika, and the authors, for the extract from the paper by K. K. Tatsuoka and M. M. Tatsuoka in Volume 52

Contents

Preface viii

1 Introduction
The aims of the book 1
The climate for language testing 2
Research and development: needs and problems 8
Research and development: an agenda 12
Overview of the book 13
Notes 15

2 Measurement in Production
Definition of terms: measurement, test, evaluation 18
Essential measurement qualities 24
Properties of measurement scales 26
Characteristics that limit measurement 30
Steps in measurement 40
Summary 49
Notes 50
Further reading 52
Discussion questions 52

3 Uses of Language Tests
Introduction 53
Uses of language tests in educational programs 53
Research uses of language tests 67
Features for classifying different types of language test 70
Summary 78
Further reading 79
Discussion questions 79

4 Communicative Language Ability
Introduction 81

Language proficiency and communicative competence 82
A theoretical framework of communicative language ability 84
Summary 107
Notes 108
Further reading 109
Discussion questions 109

5 Test Methods
Introduction 111
A framework of test method facets 116
Applications of this framework to language testing 152
Summary 156
Notes 157
Further reading 158
Discussion questions 159

6 Reliability
Introduction 160
Factors that affect language test scores 163
Classical true score measurement theory 166
Generalizability theory 187
Standard error of measurement: interpreting individual test scores within classical true score and generalizability theory 197
Item response theory 202
Reliability of criterion-referenced test scores 209
Factors that affect reliability estimates 220
Systematic measurement error 222
Summary 226
Notes 227
Further reading 232
Discussion questions 233

7 Validation
Introduction 236
Reliability and validity revisited 238
Validity as a unitary concept 241
The evidential basis of validity 243
Test bias 271
The consequential or ethical basis of validity 279
Postmortem: face validity 285

Summary 289
Notes 291
Further reading 294
Discussion questions 294

8 Some Persistent Problems and Future Directions
Introduction 296
Authentic language tests 300
Some future directions 333
A general model for explaining performance on language tests 348
Apologia et prolegomenon 351
Summary 356
Notes 358
Further reading 358
Discussion questions 358

Bibliography 361

Author index 395
Subject index 397

Preface

This book has its origins in the close personal and collaborative relationship that Buzz Palmer and I have had for a good many years. We first hatched the idea of writing a book on language testing research somewhere between the **1981** 'Illinois' study and the **1982** 'Utah' study, at a time when we were both heavily committed to trying our best to incorporate what were then the stili fairly new ideas about 'communicative competence' of people like Mike Canale, Merrill Swain, and Sandy Savignon into actual language tests, and to trying to find out if they were different from the kinds of language tests that were then most commonly used. The *two* studies that Buzz and I conducted together were a lot of hard work (neither of us may ever want to do another multitrait–multimethod study again!), but they provided a wealth of example tests and anecdotes that I have used with my classes of language testing students, and which also hopefully add a touch of both reality and comic relief to this **book.** More importantly, however, those studies forced us to face head-on some of the issues and problems that are the substance of this **book,** and to realize that addressing these will require the best ideas and tools that both applied iinguistics and psychometrics have to offer. Buzz has provided me with frequent comments and suggestions as the book has taken form and he must share the credit for the inspiration and many of the ideas herein.

Much of what is in this **book** can also be traced to two individuals whose work has influenced my research interests, and indeed my career, in very fundamental ways. My first introduction to applied linguistics was Robert Lado's (**1957)** ***Linguistics Across Cultures,*** which was required reading for ESL Peace Corps volunteers in the mid-1960s. Even though this book was quite an eye-opener for a medieval English literature major during Peace Corps training, it wasn't until I was 'in the field', teaching ESL in a high school in the Philippines, that I began to appreciate its wisdom. Its real impact on my career, however, came a few years later, when I was drawn back to it, during graduate school, after having read John **B.** Carroll's

(1964) *Language and Thought.* It was Carroll's discussions of language acquisition research and cross-cultural research in psycholinguistics, along with Lado's discussion of contrasts across languages, that I found both exciting and challenging, and that piqued an interest that eventually led me to abandon medieval literary studies for dissertation research in second language acquisition.

It was not until after graduate school, when, as a Ford Foundation 'adviser', I found myself in charge of the development and administration of language tests at a national language center in Thailand, that my on-the-job learning led me to the library, where I first discovered that either Lado or Carroll had anything to do with language testing! During the next few years I was fortunate to have the opportunity to work with John Carroll on several occasions, OR the development of language aptitude tests in Thai, and was always both awed and inspired by his encyclopedic knowledge, his brilliant insights, and his consummate craftsmanship. I continue to read his work with interest and to correspond with him on occasion to ask a question or pose a problem for his consideration. A great deal of whatever is useful in this book is a result of my contact with him and his work.

When I was trying to come up with a title for this book, it seemed that all the good titles had already been taken. There have been titles in language testing with 'issues' (for example, Oller 1983b; Alderson and Hughes 1981; Lowe and Stansfield 1988), 'current developments' (Hughes and Porter 1983), 'problems' (Upshur and Fata 1968; Interuniversidre Sprachtestgruppe Symposium Proceedings: Culhane *et al.* 1981, 1984; Klein-Braley and Stevenson 1981; Kohonen *et al.* 1985; Lutjeharms and Culhane 1982), 'approaches' (Spolsky 1978a; Brindley 1986), 'directions' (Read 1981; Lee et *al.* 1985), 'concepts' (Brière and Hinofotis 1979a) and 'research' (Oller and Perkins 1980; Oller 1983b; Bailey *et al.* 1987). And while I'm not aware of any 'principles' or 'essentials' titles in language testing, I'm not convinced that what I have to offer is quite as certain as these terms would imply. The title I've chosen turns out to be a portmanteau of the titles of two seminal works in language testing that happen to have been published in the same year: 'Fundamental considerations in the testing for English language proficiency of foreign students' (Carroll 1961a) and *Language* Testing (Lado 1961). Thus, in solving my title problem, I also echo my debt to Lado and Carroll; hopefully what I've taken from them is returned in some small measure in the pages that follow.

Throughout the travail of writing this book, I have (sometimes)

heeded the counsel, or head-bashing, if you will, of a group of individuals who have been my severest critics, and who have also aided and abetted me in this endeavor. Their written comments on various versions and parts of the manuscript have both kept me clearly attuned to fundamental issues, and pushed me to discuss areas that I might have wanted to avoid. They must therefore rightfully share the credit for what is good, and take their lumps as co-conspirators for whatever errors there are that came from them. Among those that should be thus implicated are Charles Alderson, Doug Brown, J. D. Brown, Larry Bouton, Gary Buck, Mike Canale, Gary Cziko, Fred Davidson, john de Jong, Antony Kunnan, Brian Lynch, John Oller, Sandy Savignon, Larry Selinker, Bernard Spolsky, Jack Upshur, and Swathi Vanniarajan. Comments from Gillian Brown on Chapters 4 and 5 were also very helpful. I am most grateful to Charles Alderson, John Carroll, John Clark, Bernard Spolsky, and Henry Widdowson, whose meticulous reading of the manuscript and insightful comments, from different perspectives, have improved it immensely. I would particularly like to thank Yukiko Abe-Hatasa, Buzz Palmer, Larry Selinker, and Jack Upshur for their comments and suggestions, based on their use of the book in manuscript form with their classes on language testing, and Sasi Jungsatitkul, who helped write the discussion questions. Finally, my sincerest gratitude goes to my own students, whose insights, questions, and comments have led me to sharpen my thinking on many issues, and to recognize (and admit) where I remain fuzzy and uncertain. I thank them also for patiently bearing the burden of helping me refine my presentation of these issues.

Writing this book has been challenging and rewarding in that it has given me the opportunity to work my way through some of the conundrums of language testing and to reach, if not solutions, at least a sense of direction and a strategy for research. It has also been a source of frustration, however, as I see the field moving at a pace beyond my ability to incorporate developments into the present discussion. Even as I write this preface, for example, I have received the manuscript of a 'state of the art' article on language testing from Peter Skehan, and from Liz Hamp-Lyons a review article of recent and forthcoming textbooks in applied linguistics research and language testing. These articles review recent work in language testing, and relate this to research in other areas of applied linguistics. Also in my mail is the list of titles of papers for the upcoming 11th Annual Language Testing Research Colloquium, which promise to report recent developments in a number of areas.

But while these developments may be a source of minor frustration to me, as I attempt to reach closure on this book, at the same time they give me cause for optimism. Language testers now have their own journal, ***Language Testing;*** three newsletters, ***Language Testing Update,*** the ***AILA Language Testing News,*** and the ***IATEFL Testing SIG Newsletter,*** and can count at least three major international conferences annually (the Language Testing Research Colloquium (LTRC) in North America, the Interuniversitare Sprachtestgruppe **(IUS)** Symposium in Europe, and the Academic Committee for Research on Language Testing (ACROLT) Symposium in Israel), as well as several regional conferences, such as those in Estonia, Japan, and Thailand, which regularly focus on issues in language testing. What is most encouraging about these events and developments is that the concerns of language testing are drawing together a widening circle of applied linguists, language teachers, and psychometricians, who recognize the interrelatedness of their needs, interests, and areas of expertise, and whose collaboration can only advance our understanding of language ability and how we can most effectively and usefully measure it.

Savoy, Illinois
February 1989

1 Introduction

The aims of the book

In developing and using measures of language abilities, we are constantly faced with practical questions, 'What type of test should we use?', 'How long should the test be?', 'How many tests do we need to develop?', questions to which there are no clear-cut, absolute answers. Other questions are even more difficult to answer. For example, 'How reliable should our test be?', 'Are our test scores valid for this use?', and 'How can we best interpret the results of our test?' In addressing questions such as these, we inevitably discover that the answers depend upon a wide range of prior considerations. Since these considerations will vary from one test context to the next, an appropriate answer for one situation may be inappropriate for another. Thus, in developing and using language tests we are seldom, if ever, faced with questions to which there are right or wrong answers. Answering these questions always requires consideration of the specific uses for which the test is intended, how the results are to be interpreted and used, and the conditions under which it will be given.

This book is not a 'nuts and bolts' text on how to write language tests. Rather, it is a discussion of fundamental issues that must be addressed at the start of any language testing effort, whether this involves the development of new tests or the selection of existing tests. How we conceive of these issues will affect how we interpret and use the results of language tests. One objective of this book is thus to provide a conceptual foundation for answering practical questions regarding the development and use of language tests. This foundation includes three broad areas: (1) the context that determines the uses of language tests; (2) the nature of the language abilities we want to measure, and (3) the nature of measurement. This conceptual foundation is applicable to a wide range of general concerns in language testing, including diagnostic, achievement, and language aptitude testing. Furthermore, this foundation provides a

basis for addressing issues in the measurement of language proficiency, which presents some of the most complex and challenging problems for language testing, problems to which much of the discussion of this text is addressed.

A second objective of this book is to explore some of the problems raised by what is perhaps a unique characteristic of language tests and a dilemma for language testers – that language is both the instrument and the object of measurement – and to begin to develop a conceptual framework that I believe will eventually lead, if not to their solution, at least to a better understanding of the factors that affect performance on language tests. Unlike tests of other abilities or areas of knowledge, where we frequently use language in the process of measuring something else, in language tests, we use language to measure language ability. What I believe this means is that many characteristics of the instrument, or the method of observing and measuring, will overlap with characteristics of the language abilities we want to measure. In order to understand how these characteristics interact, as I believe they do, and how they affect performance on language tests, I believe we must develop a framework for describing the characteristics of both the language abilities we want to measure and of the methods we use to measure these abilities.

The climate for language testing

Language testing almost never takes place in isolation. It is done for a particular purpose and in a specific context. A third objective of this book is thus to relate language testing to the contexts in which it takes place. Current research and development in language testing incorporates advances in several areas: research in language acquisition and language teaching, theoretical frameworks for describing language proficiency and language use, and measurement theory.'

Research in language acquisition and language teaching

As Upshur (1971) noted several years ago, there is an intrinsic reciprocal relationship between research in language acquisition and deveiopments in language teaching on the one hand, and language testing on the other. That is, language testing both serves and is served by research in language acquisition and language teaching. Language tests, for example, are frequently used as criterion measures of language abilities in second language acquisition research. Similarly, language tests can be valuable sources of

information about the effectiveness of learning and teaching. Language teachers regularly use tests to help diagnose student strengths and weaknesses, to assess student progress, and to assist in evaluating student achievement. Language tests are also frequently used as sources of information in evaluating the effectiveness of different approaches to language teaching. As sources of feedback on learning and teaching, language tests can thus provide useful input into the process of language teaching.

Conversely, insights gained from language acquisition research and ianguage teaching practice can provide valuable information for designing and developing more useful tests. For example, insights about the effects of cognitive and personality characteristics on second language acquisition have led language testers to investigate the extent to which these factors also affect performance on various types of language tests (for example, Hansen and Stansfield 1981; Stansfield and Hansen 1983; Hansen 1984; Chapelle and Roberts 1936; Chapelle 1983). And more recently, language testers have begun discussing the idea that levels of second language ability may be related to developmental sequences that characterize second language acquisition (for example, Ingram 1985; Clahsen 1985; Brindley 1986; Pienemann *et al.* 1988). Bachman (1989a) reviews areas of interface between language testing and second language acquisition research, concluding that research in areas of common concern employing a wide range of research designs and methods is likely to advance knowledge in both fields. New views of language teaching practice can also inform language test development. Mush of the development in 'communicative' language testing in the past decade, foe example (see Morrow 1977, 1979; Harrison 1983; Seaton 1983; Criper and Davies 1988; Hughes, Porter, and Weir 1988; Alderson 1988) is derived directly from the 'communicative' view of language teaching espoused by applied linguists such as Widdowson, Johnson, Brumfit, Candlin, Wilkins, and Savignon.

Thus, advances in language testing do not take place in a vacuum; they are stimulated by advances in our understanding of the processes of language acquisition and language teaching. And developments in language testing can provide both practical tools and theoretical insights for further research and development in language acquisition and language teaching.

Language ability

A clear and explicit definition of language ability is essential to all

language test development and use. Such a definition generally derives from either a language teaching syllabus or a general theory of language ability. Although much foreign/second language proficiency test development continues to be based on a skills and components framework such as those proposed by Lado (1961)and Carroll (1961a), many language testers now take a broader view of language ability. Oller, for example, has developed the notion of a 'pragmatic expectancy grammar' to characterize the abilities involved in appropriately 'mapping' aspects of discourse to the elements of the extralinguistic contexts in which language use takes place (Oller 1979b). Elsewhere, the terms 'communicative proficiency' (Bachman and Palmer 1982a), 'communicative language proficiency' (Bachman and Savignon 1986), and 'communicative language ability' (Bachman and Clark 1987; Bachman 1988) have been used to describe this broader view of language proficiency, whose distinguishing characteristic is its recognition of the importance of context beyond the sentence to the appropriate use of language. This context includes the discourse of which individual sentences are part and the sociolinguistic situation which governs, to a large extent, the nature of that discourse, in both form and function.[2]

Related to this broadened view of communicative language ability is the recognition that *communicative language use* involves a dynamic interaction between the situation, the language user, and the discourse, in which communication is something more than the simple transfer of information. This dynamic view of communication is reflected in the literature on communicative language teaching (for example, Johnson 1982; Savignon 1983) and interlanguage communication strategies (Færch and Kasper 1983). and has been included in frameworks of communicative competence (Hymes 1972b, 1982; Canale and Swain 1980; Canale 1983; Savignon 1972, 1983). This dynamic view of language use also underlies what Oller has called 'pragmatic mappings' between the elements of discourse and the extralinguistic context (Oiler 1979b).

In response to these broader views of communicative language ability and communicative language use, much effort is being directed toward developing tests that not only measure a wide range of language abilities, including grammatical, discourse, sociolinguistic, and strategic competencies, but that are also 'authentic', in that they require test takers to interact with and process both the explicit linguistic information and the implicit illocutionary or functional meaning of the test material.

A different view of language ability, which informs the Interagency

Language Roundtable (ILR) oral interview (Eowe 1982, 1985) as well as the ***ACTFL Proficiency Guidelines*** (American Council on the Teaching of Foreign Languages 1986) and the oral. interview test of language proficiency based on them, has gained considerable currency in the foreign language teaching profession. The various definitions of proficiency based on this view are derived, essentially, from the way the construct is defined operationally in the ILR and ACTFL scales. Lowe (1988), one of the major spokespersons for this view, defines proficiency as follows:

> proficiency equals achievement (ILR functions, content, accuracy) plus functional evidence of internalized strategies for creativity expressed in a single global rating of ***general language*** *ability* expressed over a wide range of functions and topics at any given ILR level. (emphasis added)
> (Lowe 1988: 12)

Lowe goes on to suggest that the two views of proficiency (the ILR/ ACTFL view and that of 'communicative language ability' outlined above) may prove incompatible, claiming that the ACTFL view is a 'holistic, top-down view', while that of communicative language ability is 'an atomistic, bottom-up view **of** language ability'. (pp. 14–15).

Proponents of the ACTFL view have claimed that the 'Guidelines' can provide a basis for criterion-referenced testing and improved professional standards (Higgs 1982b). The renewed interest in language testing that these guidelines have generated is encouraging. Nevertheless. the way in which they define language proficiency has brought to the forefront questions about the relationship between test content, test method, and the validity of interpretations or uses that are made of test scores (Bachman and Savigrion 1986; Bachman 1988a).

A common thread that runs through much recent writing in language testing **is** the belief that a precise, empirically based definition of language ability can provide the basis for developing a 'common metric' scale for measuring language abilities in a wide variety of contexts, at all levels, and in many different languages (Woodford 1978, 1981; B. J. Carroll 1980; Clark 1980; Brindley 1984). If such a scale were available, a rating of '1', for example, would always indicate the same level of ability, whether this were in listening, speaking, reading, or writing, for different contexts of language use, and even for different languages. Bachman and Clark (1987) state the advantages of a common metric as follows:

> the obvious advantage of such a scale and tests developed from it is that it would provide a standard for defining and measuring language abilities that would be independent of specific languages, contexts and domains of discourse. Scores from tests based on this scale would thus be comparable across different languages and contexts.
> (Bachman and Clark 1987: 28)

Such tests are of crucial interest for second language acquisition research and language program evaluation, where measures of language ability that can be used as criteria for comparing differences across age groups, varying native languages, and differing teaching methods are virtually nonexistent (Bachman 1989a). Such tests are equally important for use in making decisions about language competency, whether in the context of evaluating learner achievement in language programs, or for certifying the professional competence of language teachers.

Applications of measurement theory to language testing

Recently, we have seen major applications of advances in measurement theory to research and development in language testing. These applications have been primarily in four areas: construct validation, generalizability theory, item-response theory, and criterion-referenced testing.

Construct validation

Research into the relationships between performance on language tests and the abilities that underlie this performance (Construct validation research) dates at least from the 1940s, with John B. Carroll's pioneering work (Carroll 1941). The interest of language testers in the construct validity of language tests was renewed in the 1970s by John Oller's 'unitary trait hypothesis', according to which language proficiency coosists of a single unitary ability. By analyzing the relationships among scores from a wide variety of language tests, Oller believed he discovered a 'g-factor', which he interpreted as a unitary trait, 'general language proficiency'. Subsequent studies, however, disconfirmed the unitary trait hypothesis, and Oller himself eventually recognized that 'the strongest form of the unitary trait hypothesis was wrong' (Oller 1983a: *352).* Nevertheless, Oller's work, as well as the research it stimulated, firmly established construct validation as a central concern of language testing research,

and generated renewed interest in factor analysis as an analytic procedure. Other procedures have since been used to examine the construct validity of language tests, and these are discussed in greater detail in Chapter 7 below.

Generalizability theory

Generalizability theory (G-theory) provides a conceptual framework arid a set of procedures for examining several different sources of measurement error simultaneously. Using G-theory, test developers can determine the relative effects, for example, of using different test forms, of giving a test more than once, or of using different scoring procedures, and can thus estimate the reliability, or generalizability, of tests more accurately. 'G-theory' has recently been used to analyze different sources of measurement error in subjective ratings of oral interviews and writing samples, and it is discussed in detail in Chapter ***4***.

Item response theory

Item response theory (IRT) is a powerful measurement theory that provides a superior means for estimating both the ability levels of test takers and the Characteristics of test items (difficulty, discrimination). if certain specific conditions are satisfied, IRT estimates are not dependent upon specific samples, and are thus stable across different groups of individuals and across different test administrations. This makes it possible to tailor tests to individual test-takers' levels of ability, and thus to design tests that are very efficient in the **way** they measure these abilities. These characteristics are particularly useful for developing computer-adaptive tests, and item response theory **is** being used increasingly in the development and analysis of language tests. IRT also provides sample-free estimates of reliability, or precision of measurement. IRT is discussed in Chapter ***6*** below.

Criterion-referenced measurement

The measurement approach that has dominated research and development in language testing for the past twenty-five years is **that** of norm-referenced (NR)testing, in which an individual's test score **is** reported and interpreted with reference to the performance of other individuals on the test. The quintessential NR test **is** the 'standardized test' that has been tried out with large groups of individuals,

whose scores provide 'norms' or reference points for interpreting scores.

In the other major approach to measurement, that of criterion-referenced (CR) testing, test scores are reported and interpreted with reference to a specific content domain or criterion level of performance. CR tests thus provide information about an individual's mastery of a given criterion domain or ability level. While the NR approach continues to dominate the field, language testers have advocated CR measurement in some contexts, and CR principles have recently been applied to the development of language achievement tests. Furthermore, because of problems associated with the NR interpretation of test scores, the CR approach has been proposed as a basis for developing language proficiency tests for both language. program evaluation and for evaluating individual levels of ability. The CR approach is discussed more fully in Chapters 2, 6, and 8.

Research and development: needs and problems

The development and use of language tests involves an understanding, on the one hand, of the nature of communicative language use and language ability and, on the other, of measurement theory. Each of these areas is complex in its own right. Furthermore, there appear to be certain dilemmas involved in the application of current measurement models to tests that incorporate what we know about the nature of communicative language use. Language testers have thus been faced with increasingly complex problems, and have sought solutions to these problems in diverse ways.

The problems currently facing language testers have both practical and theoretical implications, and fall into two general areas. First is the problem of specifying language abilities and other factors that affect Performance on language tests precisely enough to provide a basis for test development and for the interpretation and use of test scores. The second problem is determining how scores from language tests behave as quantifications of performance. That is, what are the scaling and measurement properties of tests of language abilities? Answering this question is particularly difficult because language tests may measure several distinct but interrelated abilities. Further complications arise if we would like to interpret scores from language tests as indicators of the degree of 'mastery' with reference to some externally defined domain or criterion level of ability, rather than as indices of the relative performance of different individuals.

Defining language abilities and characterizing test authenticity

All language tests must be based on a clear definition of language abilities, whether this derives from a language teaching syllabus or a general theory of language ability, and must utilize some procedure for eliciting language performance. **As** simplistic as this statement may seem, it turns out that designing a language test is a rather complex undertaking, in which we are often attempting to measure abilities that are not very precisely defined, and using methods of elicitation that themselves depend upon the very abilities we want to measure. This is the fundamental dilemma of language testing mentioned above: the tools we use to observe language ability are themselves manifestations of language ability. Because of this, the way we define the language abilities we want to measure is inescapably related to the characteristics of the elicitation procedures, or test methods we **use** to measure these abilities. Thus, one of the most important and persistent problems in language testing is that of defining language ability in such a way that we can be sure that the test methods we use will elicit language test performance that **is** characteristic of language performance in non-test situations.

Most current frameworks of language use are based on the concept of language as communication, and recognize the importance of the context, both discourse and sociolinguistic, in which language is used. Such frameworks are based on a wealth of information from naturalistic, observational studies. I believe that there **is** now sufficient empirical evidence about the nature of language use and the abiiities that are involved in language use to begin the specification of a theoretical model of communicative language ability that will provide a basis for the development of both practical tests and of measures that can, in turn, provide the tools for the empirical investigation of this model.

A related concern has been with developing testing procedures that are 'authentic' (cf. ***Language Testing*** 2, **1,** 1985), and attempts to characterize either authenticity in general, or the authenticity of a given test have been highly problematic. Language testers have used terms such ***as*** 'pragmatic' (Oller 1979b), 'functional' (B. **3.** Carroll 1980; Farhady 1980), 'communicative' (Morrow 1979; Wesche 1981; Canale 1983), 'performance' (for example, Jones 1979b, 1985a; Courchene and de Bagheera 1985; Wesche 1985) and 'authentic' (for example, Spolsky 1985; Shohamy and Reves 1985) to characterize the extent to which the tasks required on a given test are similar to 'normal', or 'real-life' language use. However, when we consider the great variety that characterizes language use – different

contexts, purposes, topics, participants, and so forth – it is not at all clear how we might go about distinguishing 'real-life' from 'nonreal-life' language use in any meaningful way, so that attempts to characterize authenticity in terms of real-life performance are problematic. Related to this is the question of whether we can adequately reflect 'real-life language use' in language tests.

Another approach to defining authenticity in language test tasks is to adopt Widdowson's (1978) view of authentic language use as the interaction between the language user and the discourse. This notion is also implicit in Oller's (1979b) second pragmatic naturalness criterion: 'language tests. . . must require the learner to understand the pragmatic interrelationship of linguistic context and extralinguistic contexts' (Qller 1979b: 33). And while this is the approach I will advocate and expand upon in Chapter 8, it is also fraught with problems, not the least of which is the fact that different test takers are likely to interact ***individually*** in different ways with different test tasks. Some test takers, for example, may perceive a set of tasks as individual items and attempt to answer them one by one, while others may perceive them as a whole discourse, to be answered in relation to each other. Similarly, test takers may differ not only in the extent to which they are aware of and respond to the functional meaning of a given test item, but they may also have different expectations and different contexts, or what Douglas and Selinker (1985) call 'discourse domains', to which they relate that item. Since sociolinguists have been grappling with the protean nature of communicative language use in different contexts since Labov's work in the early 1970s, it is not surprising to find that variable responses to different test tasks pose a difficult problem for language testers.

Because of these problems, it is tempting simply to shrug off the question of authenticity as unimportant, as simply a matter of how the test 'appears' to the test taker. However, if authenticity is a function of the test taker's interaction with the test task, it will affect both the reliability and validity of test scores (Douglas and Selinker 1985; Qller 1986). Furthermore, the approach we take in defining authenticity is closely related to how we define language ability, and thus to how we interpret and use the results of language tests. Adequately characterizing authenticity and estimating its influence on test takers' performance is therefore one of the most pressing issues facing language testers, and constitutes a central theme of this book. I believe the key to solving this problem lies in specifying the characteristics of test tasks and test methods sufficiently well that we can begin to empirically examine test takers' performance on different types of test tasks.

Measurement concerns

A second set of problems derives from the limitations on measures of mental abilities in general, and of language abilities in particular. In this regard, we are concerned with the indirectness of our measures, the limited conditions under which we typically measure language ability, and the relatively restricted sample of performance that we obtain. Our primary concern is whether an individual's test performance can be interpreted as an indication of his competence, or ability to use language appropriately and effectively in *nontest* contexts.[3] Thus, the key measurement problem is determining the extent to which the sample of language use we obtain from a test adequately characterizes the overall potential language use of the individual. In considering this we are inevitably led to consider the question of whether the language use context of the test resembles so-called 'natural' or 'normal' nontest language use. And this, in turn, leads back to the problem of clearly describing 'natural' or 'authentic' language use.

Measurement assumptions

Our analyses and interpretations of test results are based on measurement theory, and the analytic procedures derived from this theory make specific assumptions about the nature of the abilities we test and the relationship between these abilities and scores on tests. One assumption that is fundamental to most current measurement models is that test scores are ***unidimensional,*** which means that the parts or items of a given test all measure the same, single ability. A related assumption of current measurement theory is that the items or parts of a test are ***locally independent.*** That is, we assume that an individual's response to a given test item does not depend upon how he responds to other items that are of equal difficulty.

However, from what we know about the nature of language, it is clear that virtually every instance of authentic language use involves several abilities. Listening to and comprehending a lecture, for example, requires, at least, knowledge about the sound system, lexicon and grammatical structure of the language, about the way discourse is organized, and about the sociolinguistic conventions that govern the lecturer's use of language. Furthermore, the very nature of language use is such that discourse consists of interrelated illocutionary acts expressed in a variety of related forms.

If language test scores reflect several abilities, and are thus not unidimensional, and if authentic test tasks are, by definition,

interrelated, to what extent are current measurement models appropriate for analyzing and interpreting them? The potential dilemma thus faced by language testers is that tests designed to satisfy the measurement assumptions of unidimensionality and local independence may operate at cross purposes from maintaining the authenticity of the language tasks involved, while language tests involving authentic language use, on the other hand, may be incompatible with current measurement assumptions.

The effect of test method

A final problem related to measurement theory is that of determining the extent to which test performance is influenced by the particular test method used. Numerous studies have demonstrated that test method has a sizable influence on performance on language tests (for example, Clifford 1978, 1981; Bachman and Palmer 1981a, 1982a; Shohamy 1983b, 1984a). If we are to interpret test scores as indicators of language abilities, and not of how well an individual can take multiple-choice tests, for example, we clearly need to minimize the effects of test method.

Research and development: an agenda

In addition to addressing the problems just mentioned, language testers, as applied linguists, must respond to the practical need for more appropriate measures of language abilities for use in language acquisition and language attrition research, language program evaluation, and for making decisions about individuals' attained levels of competency with respect to various educational and employment requirements. I believe that most language tests currently available are inappropriate for these purposes because they are based on a model of language ability that does not include the full range of abilities required for communicative language use, and they incorporate norm-referenced principles of test development and interpretation.

To address both the practical needs and the theoretical problems of language testing, Bachman and Clark (1987) have called for the development of a theoretical framework of factors that affect performance on language tests, and for a program of empirical research into both the measurement characteristics of language tests based on such a theoretical framework and the validity of the framework itself. This research agenda has subsequently been seconded and expanded by other language testers as well (Clark and

Clifford **1988;** Clark and Lett **1988).** Thus, one of the major themes of this book is the characterization of these factors and how their effects influence the way we interpret and use test scores. These factors fall into four categories, as illustrated in Figure 6.1 in Chapter *6* (p. **165):** communicative language ability, test method facets, personal attributes, and random factors.

The main thrust of my discussion of this theme is as follows. Some of the factors that affect scores on language tests are potentially within our control and some are not. Random factors, such as temporary fluctuations in test takers' physical condition or mental alertness, and breakdowns in equipment, are by their very nature unpredictable, and hence uncontrollable. The influence on language test performance of personal attributes, such as sex, age, native language and cultural background, background knowledge, and field independence are beginning to **be** better understood, but there are few contexts in which these can be practically controlled in the design and use of language tests. That leaves us with the characteristics, or 'facets', of the test method and communicative language ability, which, I argue, are two factors that we can and must attempt to control in the design and use of language tests. The frameworks developed in Chapters **4** and *5* of this book are presented as initial descriptions of these two sets of factors. They are also proposed as a starting place for a program of research and development which is discussed in greater detail in Chapter **8.**

The issues discussed in this book are relevant to **two** aspects of language testing: (1)the development and use of language tests; and (2) language testing research. I believe that the fundamental goal of language test development is to assure that the information, or scores, obtained from language tests will be reliable, valid and useful. This means assuring that test performance is related to and appropriate for the particular interpretations and uses for which the test is intended. I believe the fundamental goals of language testing research, on the other hand, are **(1)** to formulate and empirically validate a theory of language test performance; and (2) to demonstrate the ways in which Performance on language tests **is** related to communicative language use in its widest sense. It **is** my hope that this book will be useful for both these aspects of language testing.

Qverview of the book

Each chapter in the **book** presents a set of related issues. Following the discussion of these issues is a summary, notes, suggestions for further reading, and discussion questions.

This chapter and the next two provide a general context for the discussion of language testing. In Chapter 2 the terms 'measurement', 'test', and, 'evaluation' are defined and the relationships among them are discussed. Also described are the properties of measurement scales and the different types of measurement scales that are commonly used in language testing. Next, the essential measurement qualities of tests – reliability and validity – are introduced. Several Characteristics inherent to measures in the social sciences and the limitations these place on the interpretation and use of test scores are examined. Finally, a set of procedures for designing tests so as to minimize the effect of these limitations and maximize the reliability and validity of test scores is outlined. In Chapter 3 I discuss the various uses of language tests in educational programs, along with examples of different types of programs to illustrate these different uses. This is followed by a brief discussion of the research uses of language tests. Finally, a taxonomy for classifying different types of language tests is presented.

In Chapters 4 and 5 I present a theoretical framework for describing performance on language tests. In Chapter 4 I discuss the part of the framework that pertains to the language abilities we want to measure. 'Communicative language ability' is described as consisting of language competence, strategic competence, and psychophysiological mechanisms. In Chapter 5 I discuss the characteristics of the test methods we use to elicit language performance. These constitute facets of the testing procedure itself – the testing environment, the test rubric, the input the test taker receives, the response to that input, and the relationship between input and response. I suggest that this framework can be used both for describing the characteristics of existing language tests and for developing new language tests. I further propose that it provides a starting point for examining the reliability and validity of language tests and for formulating empirical hypotheses about the nature of performance on language tests.

Chapters 6 and 7 provide extensive discussions of the issues and problems related to demonstrating the reliability of test scores and the validity of test use. In Chapter 6 sources of error in test scores are discussed within the context of estimating the reliability of test scores. The classical true score measurement model is described and I discuss the approaches to reliability derived from it, including the assumptions, limitations and appropriate uses of these approaches. Next, some problems of the classical model are discussed. This is followed by discussions of the salient features of generaiizability

theory and item response theory as extensions of the classical model that address these problems. Next, I outline several approaches to estimating the reliability of criterion-referenced tests. I then discuss the effects of test method on test performance and how this affects our interpretation of test scores.

In Chapter 7 I discuss considerations in investigating the validity of the interpretations and uses we make of language test scores. I discuss the notion of validity as a unitary concept pertaining to a particular test interpretation or use. I then discuss the traditional approachks to validity – content, criterion, and construct – as parts of the process of validation that provide an evidential basis for the interpretation and **use** of language tests. Next, the topic of test bias is discussed, including brief discussions of some of the factors that research has shown to be potential sources of bias in language tests. Finally, I discuss validity issues related to the consequences and ethics of the use of language tests in educational systems and in society at large.

In the final chapter, I shed the mantle of objective discussant and take more of a proactive advocate's role, dealing with some persistent issues (and controversies) in language testing, and proposing an agenda for future research and development. I present what I perceive to be the pros and cons of two different approaches to defining language proficiency and authenticity in language tests, arguing that one, the 'interactional/ability' approach, provides a sounder foundation for the continued development of communicative language tests and for the validation of their use. I then argue for research and development of language tests guided by theoretical frameworks of communicarive language ability and test method facets. I further argue that such development needs to be based on criterion-referenced principles of test design and interpretation, and propose an approach to the development of criterion-referenced scales of language ability that is not based on criteria of actual language performance or actual language users. Finally, I indulge in a bit of stock-raking and crystal ball gazing, urging language testers not to lose sight of either the applied linguistic or the psychometric side of language testing, and finding both excitement at the challenges that lie ahead and confidence in our ability po meet them.

Notes

1 Although many researchers distinguish language learning from language acquisition, I will **use** the term 'language acquisition'

in a nontechnical sense throughout this book to refer to the process of attaining the ability to use language.

2 The term 'language proficiency' has been traditionally used in the context of language testing to refer in general to knowledge, competence, or ability in the use of a language, irrespective of how, where, or under what conditions it has been acquired (for example, Carroll 1961a; Davies 1968b; Spolsky 1968; Upshur 1979; Oller 1979b; Rivera 1984). Another term that has entered the context of language testing, from linguistics via language teaching, is 'communicative competence', which also refers to ability in language use, albeit a broader view of such use than has been traditionally associated with the term 'language proficiency' (for example, Hymes 1972b; Savignon 1972, 1983; Canale and Swain 1980). Recently, the term 'proficiency' has come to be associated, in foreign language teaching circles, almost exclusively with a specific language testing procedure, the ACTFWILR Oral Proficiency Interview (Lowe 1983, 1985; Liskin-Gasparro 1984; American Council on the Teaching of Foreign Languages 1986).

The term 'proficiency' has thus acquired a variety of meanings and connotations in different contexts. Therefore, in order to forestall misinterpretation and, if possible, to **facilitate** the discussion of issues of concern to language testing, I want to clarify the usage that will be followed in this book. The term I prefer to use is simply 'language ability'. However, at times it **is** necessary to use the term 'language proficiency', and in such cases in this book it is essentially synonymous with 'language ability', or ability in language use.

3 It has become common practice to offer some sort of stylistic solution to the problems related to writing in a language which no longer has a neuter gender in its singular personal pronouns. One solution that I have decided against is the use of 'he or she' or 's/he', since this commits an almost equally grave infelicity, in my opinion, of dehumanizing the language. Another solution, particularly popular among male writers, it seems, is to offer a blanket disclaimer of sexism in language, and to then somehow justify the use of the masculine forms of pronouns on the basis of stylistic consistency. I find this approach personally unattractive, since it is inconsistent with my own beliefs about sexism in general. The approach I will use is to alternate between masculine and feminine forms (except, of course, when referring to specific persons whose sex is known). But rather than

accomplishing this alternation more or less at random, as happens in the human population, I will impose a sort of *systematicity* to this alternation, maintaining a given gender or combination of genders throughout a thematic section or extended example, and then switching this in the following section. I will, of course, make every attempt to avoid any sexual stereotyping.

2 Measurement

Introduction

In developing language tests, we must take into account considerations and follow procedures that are characteristic of tests and measurement in the social sciences in general. Likewise, our interpretation and use of the results of language tests are subject to the same general limitations that characterize measurement in the social sciences. The purpose of this chapter is to introduce the fundamental concepts of measurement, an understanding of which is essential to the development and use of language tests. These include the terms 'measurement', 'test', and 'evaluation', and how these are distinct from each other, different types of measurement scales and their properties, the essential qualities of measures – reliability and validity, and the characteristics of measures that limit our interpretations of test results. The process of measurement is described as a set of steps which, if followed in test development, will provide the basis for both reliable test scores and valid test use.

Definition of terms: measurement, test, evaluation

The terms 'measurement', 'test', and 'evaluation' are often used synonymously; indeed they may, in practice, refer to the same activity.' When we ask for an evaluation of an individual's language proficiency, for example, we are frequently given a test score. This attention to the superficial similarities among these terms, however, tends to obscure the distinctive characteristics of each, and I believe that an understanding of the distinctions among the terms is vital to the proper development and use of language tests.

Measurement

Measurement in the social sciences is the process of quantifying the characteristics of persons according to explicit procedures and rules.[2]

This definition includes three distinguishing features: quantification, characteristics, and explicit rules and procedures.

Quantification

Quantification involves the assigning of numbers, and this distinguishes measures from qualitative descriptions such as verbal accounts or nonverbal, visual representations. Non-numerical categories or rankings such as letter grades (**'A,** B, **C ...**'), or labels (for example, 'excellent, good, average ...') may have the characteristics of measurement, and these are discussed below under 'properties of measurement scales' (pp. **26–30).** However, when we actually use categories or rankings such as these, we frequently assign numbers to them in order to analyze and interpret them, and technically, it is not until we do this that they constitute measurement.

Characteristics

We can assign numbers to both physical and mental characteristics **of** persons. Physical attributes such as height and weight can be observed directly. In testing, however, we are almost always interested in quantifying mental attributes and abilities, sometimes called traits or constructs, which can only be observed indirectly. These mental attributes include characteristics such as aptitude, intelligence, motivation, field dependence/independence, attitude, native language, fluency in speaking, and achievement in reading comprehension.

The precise definition of 'ability' is a complex undertaking. In a very general sense, 'ability' refers to being able to do something, but the circularity of this general definition provides little **help** for measurement unless we can clarify what the 'something' is. John B. Carroll (1983c, 1987a) has proposed defining an ability with respect to a particular class of cognitive or mental tasks that an individual is required to perform, and 'menta1 ability' thus refers to performance on a set of mental tasks (Carroll 1987a: **268).** We generally assume that there are degrees of ability and that these are associated with tasks or performances of increasing difficulty or complexity (Carroll 1980, 1987a). Thus, individuals with higher degrees of a given ability could be expected to have a higher probability of correct performance on tasks of lower difficulty or complexity, and a lower probability of correct performance on tasks of greater difficulty or complexity.

Whatever attributes or abilities we measure, it is important to understand that it is these attributes or abilities and ***not*** the persons themselves that we are measuring. That is, we are far from being able to claim that a single measure or even a battery of measures can adequately characterize individual human beings in all their complexity.

Rules and procedures

The third distinguishing characteristic of measurement is that quantification must be done according to explicit rules and procedures. That is, the 'blind' or haphazard assignment of numbers to characteristics of individuals cannot be regarded as measurement. In order to be considered a measure, an observation of an attribute must be replicable, for other observers, in other contexts and with other individuals. Practically anyone can rate another person's speaking ability, for example. But while one rater may focus on pronunciation accuracy, another may find vocabulary to be the most salient feature. Or one rater may assign a rating as a percentage, while another might rate on a scale from zero to five. Ratings such as these can hardly be considered anything more than numerical summaries of the raters' personal conceptualizations of the individual's speaking ability. This is because the different raters in this case did not follow the same criteria **or** procedures for arriving at their ratings. Measures, then, are distinguished from such 'pseudo-measures' by the explicit procedures and rules upon which they are based. There are many different types of measures in the social sciences, including rankings, rating scales, and tests.[3]

Test

Carroll (1**968)** provides the following definition of a test:

> a psychological or educational test is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual.
> (Carroll **1968: 46)**

From this definition, it follows that a test is a measurement instrument designed to elicit a specific sample of an individual's behavior. **AS** one type of measurement, a test necessarily quantifies characteristics of individuals according to explicit procedures. What distinguishes a test from other types of measurement is that it is

designed to obtain a specific sample of behavior. Consider the following example. The Interagency Language Roundtable (ILR) oral interview (Lowe 1982), is a test of speaking consisting of (1) a set of elicitation procedures, including a sequence of activities and sets of question types and topics; and (2) a measurement scale of language proficiency ranging from a low level of '0' to a high level of '5', on which samples of oral language obtained via the elicitation procedures are rated. Each of the six scale levels is carefully defined by an extensive verbal description. A qualified ILR interviewer might be able to rate an individual's oral proficiency in a given language according to the IER rating scale, on the basis of several years' informal contact with that individual, and this could constitute a measure of that individual's oral proficiency. This measure could not be considered a test, however, because the rater did not follow the procedures prescribed by the ILR oral interview, and consequently may not have based her ratings on the kinds of specific language performance that are obtained in conducting an ILR oral interview.

I believe this distinction is an important one, since it reflects the primary justification for the use of language tests and has implications for how we design, develop, and use them. If we could count on being able to measure a given aspect of language ability on the basis of *any* sample of language use, however obtained, there would be no need to design language tests. However, it is precisely because any given sample of language will not necessarily enable the test user to make inferences about a **given** ability that we need language tests. That is, the inferences and uses we make of language test scores depend upon the sample of language use obtained. Language tests can thus provide the means for more carefully focusing on the specific language abilities that are of interest. As such, they could be viewed as supplemental to other methods of measurement. Given the limitations on measurement discussed below (pp. 30–40), and the potentially large effect of elicitation procedures on test performance, however, language tests can more appropriately be viewed as the best means of assuring that the sample of language obtained is sufficient for the intended measurement purposes, even if we are interested in very general or global abilities. That is, carefully designed elicitation procedures such as those of the ILR oral interview, those for measuring writing ability described by Jacobs *et al.* (1981), or those of multiple-choice tests such as the ***Test of English as a Foreign Language*** (TOEFL), provide the best assurance that scores from language tests will be reliable, meaningful, and useful.[4]

While measurement is frequently based on the naturalistic observation of behavior over a period of time, such as in teacher rankings or grades, such naturalistic observations might not include samples of behavior that manifest specific abilities or attributes. Thus a rating based on a collection of personal letters, for example, might not provide any indication of an individual's ability to write effective argumentative editorials for a news magazine. Likewise, a teacher's rating of a student's language ability based on informal interactive social language use may nor be a very good indicator of how well that student can use language to perform various 'cognitive/ academic' language functions (Curnmins 1980a). This is not ob imply that other measures are less valuable than tests, but to make the point that the value of tests lies in their capability for eliciting the specific kinds of behavior that the test user can interpret as evidence of the attributes or abilities which are of interest.

Evaluation

Evaluation can be defined as the systematic gathering of information for the purpose of making decisions (Weiss 1972).[5] The probability of making the correct decision in any given situation is a function not only of the ability of the decision maker, but also of the quality of the information upon which the decision is based. Everything else being equal, the more reliable and relevant the information, the better the likelihood of making the correct decision. Few of us, for example, would base educational decisions on hearsay or rumor, since we would not generally consider these to be reliable sources of information. Similarly, we frequently attempt to screen out information, such as sex and ethnicity, that we believe **to** be irrelevaat to a particular decision. One aspect of evaluation, therefore, is the collection of reliable and relevant information. This information need not be, indeed seldom is, exclusively quantitative. Verbal descriptions, ranging from performance profiles to letters of reference, as well as overall impressions, can provide important information for evaluating individuals, **as** can measures, such as ratings and test scores.

Evaluation, therefore, does not necessarily entail testing. By the same token, tests in and of themselves are not evaluative. Tests are often used for pedagogical purposes, either as a means of motivating students to study, or as a means of reviewing material taught, in which case no evaluative decision is made on the basis of the test results. Tests may also be used for purely descriptive purposes. It is

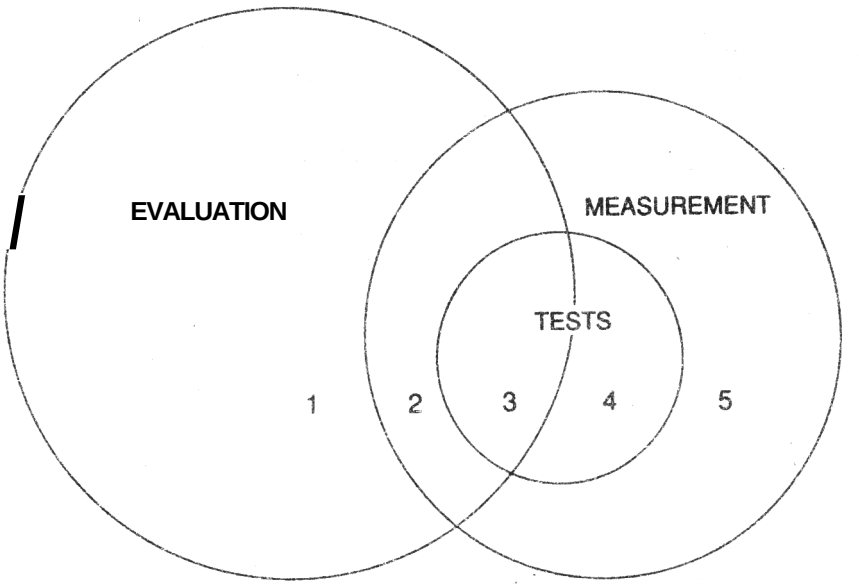

Figure* 2.1** ***Relationships among measurement, tests, and evaluation

only when the results of tests are used as a basis for making a decision that evaluation is involved. Again, this may seem a moot point, but it places the burden for much of the stigma that surrounds testing squarely upon the test user, rather than on the test itself. Since by far the majority of tests are used for the purpose of making decisions about individuals, I believe it is important to distinguish the information-providing function of measurement from the decision-making function of evaluation.

The relationships among measurement, tests, and evaluation are illustrated in Figure 2.1. An example of evaluation that does not involve either tests or measures (area '1') is the use of qualitative descriptions of student performance for diagnosing learning problems. An example of a ***non-test measure*** for evaiuasion (area '2') is a teacher ranking used for assigning grades, while an example of a ***test*** used 'for purposes of evaluation (area '3') is she use of an achievement test to determine student progress. The most common non-evaluative uses of tests and measures are for research purposes. ***An*** example of tests that are not used for evaluation (area **'4')** is the use **of** a proficiency test as a criterion in second language acquisition research. Finally, assigning code numbers to subjects in second language research according to native language is an example **of** a *non-test*

measure that is not used for evaluation (area **'5').** In summary, then, not all measures are tests, not all tests are evaluative, and not all evaluation involves either measurement or tests.

Essential measurement qualities

If we are to interpret the score on a given test as an indicator of an individual's ability, that score must be both reliable and valid. These qualities are thus essential to the interpretation and use of measures of language abilities, and they are the primary qualities to be considered in developing and using tests.

Reliability

Reliability is a quality of test scores, and a perfectly reliable score, or measure, would be one which is free from errors of measurement (American Psychological Association 1985). There are many factors other than the ability being measured that can affect performance on tests, and that constitute sources of measurement error. Individuals' performance may be affected by differences in testing conditions, fatigue, and anxiety, and they may thus obtain scores that are inconsistent from one occasion to the next. If, for example, a student receives a low score on a test one day and a high score on the same test two days later, the test does not yield consistent results, and the scores cannot be considered reliable indicators of the individual's ability. Or suppose two raters gave widely different ratings to the same writing sample. In the absence of any other information, we have no basis for deciding which rating to use, and consequently may regard both as unreliable. Reliability thus has to **do** with the consistency of measures across different times, test forms, raters, and other characteristics of the measurement context.

In any testing situation, there are likely to be severai different sources of measurement error, so that the primary concerns in examining the reliability of test scores are first, to identify the different sources of error, and then to use the appropriate empirical procedures for estimating the effect of these sources of error on test scores. The identification of potential sources of error involves making judgments based **on** an adequate theory of sources of error. Determining how much these sources of error affect test scores, on the other hand, is a matter of empirical research. The different approaches to defining and empirically investigating reliability will be discussed in detail in Chapter 6.

Validity

The most important quality of test interpretation or use is validity, or the extent to which the inferences or decisions we make on the basis of test scores are ***meaningful, appropriate,*** and ***useful*** (American Psychological Association 1985). In order for a test score to be a meaningful indicator of a particular individual's ability, we must be sure it measures that ability and very little else. Thus, in examining the meaningfulness of test scores, we are concerned with demonstrating that they are not unduly affected by factors other than the ability being tested. If test scores are strongly affected by errors of measurement, they will not be meaningful, and cannot, therefore, provide the basis for valid interpretation or use. A test score that is not reliable, therefore, cannot be valid. If test scores are affected by abilities other than the one we want to measure, they will not be meaningful indicators of that particular ability. If, for example, we ask students to listen to a lecture and then to write a short essay based on that lecture, the essays they write will be affected by both their writing ability and their ability to comprehend the lecture. Ratings of their essays, therefore, might not be valid measures of their writing ability.

In examining validity, we must also be concerned with the appropriateness and usefulness of the test score for a given purpose. A score derived from a test developed to measure the language abilities of monolingual elementary school children, for example, might not be appropriate for determining the second language proficiency of bilingual children of the same ages and grade levels. To use such a test for this latter purpose, therefore, would be highly questionable (and potentially illegal). Similarly, scores from a test designed to provide information about an individual's vocabulary knowledge might not be particularly useful for placing students in a writing program.

While reliability is a quality of test scores themselves, validity is a quality of test Interpretation and use. As with reliability, the investigation of validity is both a matter of judgment and of empirical research, and involves gathering evidence and appraising the values and social consequences that justify specific interpretations or uses of test scores. There are many types of evidence that can be presenred to support the validity of a given test interpretation or use, and hence many ways of investigaring validity. Different types of evidence that are relevant to the investigation of validity and approaches to collecting this evidence are discussed in Chapter 7.

Reliability and validity are both essential to the use of tests.

Neither, however, is a quality of tests themselves; reliability is a quality of test scores, while validity is a quality of the interpretations or uses that are made of test scores. Furthermore, neither is absolute, in that we can never attain perfectly error-free measures in actual practice, and the appropriateness of a particular use of a test score will depend upon many factors outside the test itself. Determining what degree of relative reliability or validity. is required for a particular test context thus involves a value judgment on the part of the test user.

Properties of measurement scales

If we want to measure an attribute or ability of an individual, we need to determine what set of numbers will provide the best measurement. When we measure the loudness of someone's voice, for example, we use decibels, but when we measure temperature, we use degrees Centigrade or Fahrenheit. The sets of numbers used for measurement must be appropriate to the ability or attribute measured, and the different ways of organizing these sets of numbers constitute ***scales of measurement.***

Unlike physical attributes, such as height, weight, voice pitch, and temperature, we cannot directly observe intrinsic attributes or abilities, and we therefore must establish our measurement scales by definition, rather than by direct comparison. The scales we define can be distinguished in terms of four properties. A measure has the property of ***distinctiveness*** if different numbers are assigned to persons with different values on the attribute, and is ***ordered in magnitude*** if larger numbers indicate larger amounts of the attribute. If equal differences between ability levels are indicated by equal differences in numbers, the measure has equal ***intervals,*** and if a value of zero indicates the absence of the attribute, the measure has an ***absolute zero point.***

Ideally, we would like the scales we use to have all these properties, since each property represents a different type of information, and the more information our scale includes, the more useful it will be for measurement. However, because of the nature of the abilities we wish to measure, as well as the limitations on defining and observing the behavior that we believe to be indicative of those abilities, we are not able to use scales that possess all four properties for measuring every ability. That is, not every attribute we want to measure, or quantify, fits on the same scale, and not every procedure we use for observing and quantifying behavior yields the same scale, so that it is

necessary to use different scales of measurement, according to the characteristics of the attribute we wish to measure and the type of measurement procedure we use. Ratings, for example, might be considered the most appropriate way to quantify observations of speech from an oral interview, while we might believe that the number of items answered correctly on a multiple-choice test is the best way to measure knowledge of grammar. These abilities are different, as are the measurement procedures used, and consequently, the scales they yield have different properties. The way we interpret and use scores from our measures is determined, to a large extent, by the properties that characterize the measurement scales we use, and it is thus essential for both the development and the use of language tests to understand these properties and the different measurement scales they define. Measurement specialists have defined four types of measurement scales – ***nominal, ordinal, interval,*** and ***ratio*** – according to how many of these four properties they possess.'

Nominal **scale**

As its name suggests, a nominal scale comprises numbers that are used to 'name' the classes or categories of a given attribute. That is, we can use numbers as a shorthand code for identifying different categories. If we quantified the attribute 'native language', for example, we would have a nominal scale. We could assign different code numbers to individuals with different native language backgrounds, (for example, Amharic = 1, Arabic = 2, Rengali = 3, Chinese = 4, etc.) and thus create a nominal scale for this attribute. The numbers we assign are arbitrary, since it makes no difference what number we assign to what category, so long as each category has a unique number. The distinguishing characteristic of a nominal scale is that while the categories to which we assign numbers are distinct, they are *not* ***ordered*** with respect to each other. In the example above, although '1' (Amharic) is ***not equal to*** '2' (Arabic), it is neither greater than nor less than '2'. Nominal scales thus possess the property of distinctiveness. Because they quantify categories, nominal scales are also sometimes referred to as 'categorical' scales. A special case of a nominal scale is a ***dichotomous scale,*** in which the attribute has only two categories, such as 'sex' (male and female), or 'status of answer' (right and wrong) on some types of tests.

OXFORD APPLIED LINGUISTICS

Fundamental Considerations in Language Testing

This book explores the basic considerations that underlie the practical development and use of language tests: the nature of measurement, the contexts that determine the use of language tests, and the nature of both the language abilities to be measured and the testing methods that are used to measure them. It also serves as a synthesis of much of the research that hitherto has only been available in collections of readings.

Students on teacher education courses and all those professionally involved in both the development and use of language tests will find it an authoritative and illuminating complement to practical 'how to' books.

Lyle F. Bachman is Professor of Applied Linguistics at the University of California, Los Angeles and Professor and Chair of English Language Teaching at The Chinese University of Hong Kong.

Reviews of *Fundamental Considerations in Language Testing:*

'There is much of value in Bachman's book for practising teachers and students in TESOL courses.'

Liz Hamp-Lyons, *TESOL Quarterly*

'*Fundamental Considerations in Language Testing* is an extremely valuable addition to the literature on the measurement of language skills.'

Daniel L. Robertson, *TESOL Journal*

Oxford University Press
ISBN 0-19-437003-8