

TABLE OF CONTENTS

Preface	x
---------------	---

Chapter 1

Types and Uses of Language Tests	1
Two Families of Language Tests	1
Norm-Referenced Tests	2
Criterion-Referenced Tests	2
Type of Interpretation	3
Type of Measurement	4
Purpose of the Testing	5
Distributions of Scores	5
Test Structure	5
Matching Tests to Decision Purposes	7
Program-Level Proficiency Decisions	8
Program-Level Placement Decisions	9
Classroom-Level Achievement Decisions	11
Classroom-Level Diagnostic Decisions	12
Why a Single Test Cannot Fulfill All Four Functions	12
Differences in Ranges of Ability	13
Differences in Variety of Content	14
Using Spreadsheet Programs in Language Testing	15
What Is a Spreadsheet Program?	16
How Will You Personally Benefit from Using a Spreadsheet Program in This Book?	16
Review Questions	17
Application Exercises	17

Chapter 2

Adopting, Adapting, and Developing Language Tests	18
Theoretical Issues	18
Language Teaching Methodology Issues	19
An Exceptionally Short History of Language Testing	19
Why Knowing about These Movements Is Important	24
The Competence/Performance Issue	24
The Discrete-Point/Integrative Issue	25
• Practical Issues	26
The Fairness Issue	26
The Cost Issues	27
Ease of Test Construction	28
Ease of Test Administration	28
Ease of Test Scoring	29
Interactions of Theoretical Issues	29
Adopt, Adapt, or Develop?	30
Adopting Language Tests	30

Adapting Language Tests	33
Developing Language Tests	34
Putting Sound Tests in Place	34
Getting Started with Your Spreadsheet Program	36
Moving Around the Spreadsheet	37
Creating a Sample Spreadsheet	38
Entering Test Score Data to Create a Spreadsheet	38
Review Questions	40
Application Exercises	40

Chapter 3	Developing Good Quality Language Test Items	41
	What is a Test Item?	41
	Guidelines for Item Format Analysis	42
	General Guidelines	43
	Receptive Response Items	47
	Productive Response Items	51
	Personal Response Items	58
	Why Bother with Item Format Analysis?	63
	Review Questions	64
	Application Exercises	64

Chapter 4	Item Analysis in Language Testing	66
	Norm-Referenced Item Analysis	66
	Item Facility Analysis	66
	Item Discrimination Analysis	68
	Calculating Item Facility and Discrimination with Your Spreadsheet	70
	NRT Development and Improvement Projects	75
	Criterion-Referenced Item Analysis	76
	Item Quality Analysis	77
	CRT Development and Improvement Projects	79
	Role of Item Facility	80
	Difference Index	80
	The B-Index	82
	CRT Item Selection	84
	Review Questions	85
	Application Exercises	86

Chapter 5	Describing Language Test Results	89
	Displaying Data	89
	Graphic Display of Frequencies	91
	Creating Graphs in Excel™	93
	Scales of Measurement	95

Nominal Scales	95
Ordinal Scales	96
Continuous Scales	96
Descriptive Statistics	97
Central Tendency	98
Mean	98
Mode	99
Median	100
Midpoint	100
Dispersion	101
Range	101
High and Low	102
Standard Deviation	102
Variance	104
The Spreadsheet Approach to Descriptive Statistics	105
Reporting Descriptive Statistics	107
What Should Be Included?	107
How Should Descriptive Test Statistics be Displayed?	108
Review Questions	110
Application Exercises	111

Chapter 6 Interpreting Language Test Scores..... 114

Probability Distributions	114
Normal Distribution	116
Characteristics of Normal Distributions	119
Central Tendency	119
Dispersion	119
Percents/Percentages	120
Learning from Distributions	121
Using Percents/Percentages	122
Percentiles	122
Standardized Scores	123
z Scores	123
T Scores	125
CEEB Scores	126
Computer-based TOEFL Scores	126
Standardized and Percentile Scores	126
The Importance of Standardized Scores	127
Skewed Distributions	129
Skewedness	129

Peaked Distributions	132
NRT and CRT Distributions	132
The Spreadsheet Approach to Standardized Scores	134
Review Questions	136
Application Exercises	137

Chapter 7 Correlation in Language Testing..... 139

Preliminary Definitions.....	139
Calculating the Pearson Product-Moment Correlation Coefficient	142
Assumptions of the Pearson-Moment Correlation Coefficient	145
Calculating the Pearson Correlation Coefficient with a Spreadsheet	149
Interpreting Correlation Coefficients	153
Statistical Significance	153
Meaningfulness	157
Correlation Matrixes	159
Potential Problems with Correlational Analysis.....	161
Restriction of Range	161
Skewedness	161
Causality	162
Another Useful Type of Correlational Analysis	162
Point-Biserial Correlation Coefficient	162
Calculating the Point-Biserial Correlation Coefficient with a Spreadsheet.....	164
Review Questions	167
Application Exercises	168

Chapter 8 Language Test Reliability 169

Sources of Variance	169
Measurement Error	171
Variance Due to Environment	172
Variance Due to Administration Procedures	173
Variance Attributable to Examinees	173
Variance Due to Scoring Procedures	174
Variance Attributable to the Test and Test Items	174
Reliability of NRTs	175
Test-Retest Reliability	175
Equivalent-Forms Reliability	176
Internal-Consistency Reliability	176
Split-Half Reliability.....	177
Cronbach Alpha	179
Kuder-Richardson Formulas	179

Reliability of Rater Judgments	185
Interpreting Reliability Estimates	188
Standard Error of Measurement	188
Using a Spreadsheet to Calculate NRT Reliability	190
Split-Half (Adjusted) and Cronbach Alpha	190
Kuder-Richardson Formulas 20 and 21, and the SEM	193
Factors Affecting the Reliability of NRTs	196
Review Questions	196
Application Exercises	197

Chapter 9 Language Test Dependability 199

Threshold Loss Agreement Approaches	200
Agreement Coefficient	200
Kappa Coefficient	202
Estimating Threshold Loss Agreement from a Single Test Administration	203
Squared-Error Loss Agreement Approaches	206
Domain Score Dependability	207
Relationships	209
Cautions	209
Confidence Intervals	210
Using a Spreadsheet to Calculate CRT Reliability	210
Agreement and Kappa Estimates (Using the Subkoviak Method)	211
The Phi(λ) Coefficient	211
The Phi Coefficient and Confidence Interval	214
Factors Affecting the Consistency of CRTs	215
Review Questions	216
Application Exercises	217

Chapter 10 Language Test Validity 220

Traditional Validity Strategies of Both NRTs and CRTs	221
Content Validity	221
Overall Strategy for Establishing Content Validity	221
An Example of the Importance of Item Planning in Regards to Content Validity	225
Content Validity and Other Types of Validity	226
Construct Validity	226
Differential-Groups Studies	227
Intervention Studies	232
Criterion-Related Validity: A Traditional Strategy for NRTs	233
Restrictions of Range and NRT Validity	234
Standard Setting	235
Reliability, Validity, and Standard Setting	238

Standards and Test Consistency	238
Standards and Test Validity	239
Other Issues Related to Validity	242
The Washback Effect	242
Factors Affecting the Impact of Washback	243
Negative Effects of Washback	243
Promoting Positive Washback	244
Testing Bias	246
Review Questions	248
Application Exercises	249

Chapter 11	Language Testing in Reality	252
	The Place of Tests in Curriculum Planning	252
	The ELI Language Program	254
	Four Decision-Making Steps	254
	Initial Screening Procedures	254
	Placement Procedures	256
	Second-Week Diagnostic Procedures	257
	Achievement Procedures	258
	Testing as an Integrated System	260
	Review Questions	261
	Application Exercises	261
	Answer Keys	262
	Glossary	284
	Index	296
	References	303

PREFACE

As is often true in the language teaching field, this volume had its roots in a class that I teach quite regularly—in this case, a graduate-level course in language testing. While many books exist on language testing, none seemed to offer the types of information that I wanted to present in my class. I felt that some books were too technical and complex to be thoroughly covered in one semester, while others were too practical—offering many ideas for different types of language test questions, but very little on test construction, analysis, and improvement. As a result, this language testing book is designed to cover the middle ground. I have tried to provide a balance between the technical and practical aspects of language testing that is neither too complex nor too simplistic.

My overall goal was to provide information about language testing that would not only be immediately useful for making program-level decisions (e.g., admissions and placement decisions), but also information about testing for classroom-level decisions (i.e., assessing what the students have learned through diagnostic or achievement testing). These two categories of decisions and the types of tests that are typically used to make them are quite different.

The category of tests most useful for program-level decisions consists of tests specifically designed to compare the performances of students to each other. These are called norm-referenced tests because interpretation of the scores from this category of tests is linked closely to the notion of the normal curve (also known as the “bell” curve). Such tests are most commonly used to spread students out along a continuum of scores based on some general knowledge or skill area so that the students can be placed, or grouped, into ability levels. The administrator's goal in using this type of test is usually to group students of similar abilities together in order to make the teacher's job easier. In other situations, the administrator may be interested in making comparisons between the average proficiency levels of students in different levels, between different language institutions or among students across the nation. Norm-referenced tests are also appropriate for language proficiency testing. Notice that the purpose of the tests in the norm-referenced family is to make comparisons in performance either between students within an institution (for placement purposes) or between students across courses or institutions (for proficiency assessment purposes). In short, sound norm-referenced tests can help administrators and teachers do their jobs better.

In contrast, the criterion-referenced family of tests is most useful to teachers in the classroom (though administrators should be interested in these tests as well). Criterion-referenced tests are specifically designed to assess how much of the material or set of skills taught in a course is being learned by the students. With criterion-referenced tests, the purpose is not to compare the performances of students to each other, but rather to look at the performance of each individual student vis-à-vis the material or curriculum at hand. They are called criterion-referenced tests because interpretation of the scores is intimately linked to assessing well-defined criteria for what is being taught. Such tests are often used to diagnose the strengths and weaknesses of students with regard to the goals and objectives of a course or program. At other times, criterion-referenced tests may be used to assess achievement, in the sense of “how much has each student learned.” Such information may be useful for grading student performance in the course, or for deciding whether to promote the students to the next level of study, as well as for improving the materials, presentation, and sequencing of teaching points. In short, sound criterion-referenced tests can help the teacher do a better job.

My primary motivation in writing this book was to provide practical and useful testing tools that will help language program administrators and teachers do their respective jobs better. The distinction between the norm-referenced and criterion-referenced tests will help administrators and teachers focus on the respective types of tests most appropriate for the kinds of decisions that they make in their work. Hence the topic of each chapter will be approached from both norm-referenced and criterion-referenced perspectives. After all, the decisions made by administrators and teachers affect students' lives, sometimes in dramatic ways, involving a great deal of time and money, other times in more subtle ways, including psychological and attitudinal factors.

I assume that teachers, though most interested in classroom tests, will also take an interest in program-level decisions. Similarly, I assume that administrators, though primarily interested in program-level decisions, will also take an interest in classroom-level tests. Each group is inevitably involved in the other's decision making—perhaps in the form of teachers proctoring and scoring the placement test, or perhaps in the form of an administrator evaluating the effectiveness of teachers' classroom tests. The types of decisions discussed in this book may interact in innumerable ways, and I think that any cooperation between administrators and teachers in making decisions will be healthy for the curriculum in general and test development in particular.

Regardless of whether the reader is a teacher, an administrator, or both, the goal of reading this book should be to learn how to do all types of testing well. Inferior or mediocre testing is common, yet most language professionals recognize that such practices are irresponsible and eventually lead to inferior or mediocre decisions being made about their students' lives. The tools necessary to do high quality testing are provided in this book. Where statistics are involved, they are explained in a straightforward "recipe book" style so that readers can immediately understand and apply what they learn to their teaching or administrative situations. If this book makes a difference in the quality of decision making in even one language program, the time and effort that went into writing it will all have been worthwhile.

This is the second edition of this book. Brown (1996a) was the first edition, and Brown (translated by Wada 1999) provided a Japanese translation. This edition differs in several ways from the first edition. Most prominently, this edition has been updated throughout to reflect the present state of knowledge on all the topics covered, including many new sections and new references. But also of importance, based on the feedback and suggestions of professors using the first edition of the book, the conceptual and computational explanations of the various statistical techniques in the first edition have been expanded to include clear directions for doing the various statistics in a spreadsheet computer program. Judging by feedback from readers, the first edition of this book was found to be useful by many. I hope this new expanded edition will prove even more useful in real language teaching situations like yours.

I would like to thank Kathleen Bailey, John Nelson, and Betsy Parrish for their helpful comments during the reviewing process. Also, I would like to thank Mark Nelson and Sophia Wisener for their help in the editing process.

Finally, I would like to thank Microsoft for permission to use their *Excel*TM program.