TABLE OF CONTENTS

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 proccess. 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*™ program.