Preface

The publication of Spearman's paper "The proof and measurement of association between two things" in the *American Journal of Psychology* in 1904 was the very tentative start of a new field now known as test theory. This book appears almost exactly a century later. During this period, test theory has developed from a timid fledgling to a mature discipline, with numerous results that nowadays support item and test analysis and test scoring at nearly every testing organization around the world.

This preface is not an appropriate place to evaluate a hundred years of test theory. But two observations may help me to explain my motives for writing this book. The first is that test theory has developed by careful modeling of response processes on test items and by using sophisticated statistical tools for estimating model parameters and evaluating model fit. In doing so, it has reached a current level of perfection that no one ever thought possible, say, two or three decades ago. Second, in spite of its enormous progress, although test theory is omnipresent, its results are used in a peculiar way. Any outsider entering the testing industry would expect to find a spin-off in the form of a well-developed technology that enables us to engineer tests rigorously to our specifications. Instead, test theory is mainly used for post hoc quality control, to weed out unsuccessful items, sometimes after they have been pretested, but sometimes after they have already been in operational use. Apparently, our primary mode of operation is not to create good tests, but only to prevent bad tests. To draw a parallel with the natural sciences, it seems as if testing has led to the development of a new science, but the spin-off in the form of a technology for engineering the test has not yet been realized.

Part of the explanation for our lack of technology may be a deeply ingrained belief among some in the industry that test items are unique and that test development should be treated as an art rather than a technology. I certainly believe that test items are unique. In fact, I even hope they will remain so; testing would suffer from serious security problems if they ceased to be so. Also, as a friend of the arts, I am sensitive to the aesthetic dimension of human artifacts. The point is, however, that these qualities do not relieve testing professionals of their duty to develop a technology. To draw another parallel, architecture has a deep artistic quality to it, and good architects are true artists. But if they were to give up their technology, we would have no place to live or work.

The use of design principles is an essential difference between technology-based approaches and the approaches with post hoc quality control hinted at above. Another difference is the use of techniques to guarantee that products will operate according to our specifications. These principles and techniques are to be used in a process that goes through four different stages: (1) establishing a set of specifications for the new testing program, (2) designing an item pool to support the program, (3) developing the item pool, and (4) assembling tests from the pool to meet the specifications. Although it is essential that the first stage be completed before the others are, the three other stages are more continuous and are typically planned to optimize the use of the resources in the testing organization. But it is important to distinguish between them because each involves the use of different principles and techniques.

At a slightly more formal level, test design is not unique at all; some of its stages have much in common with entirely different areas, where professionals also develop products, have certain goals in mind, struggle with constraints, and want optimal results. In fact, in this book I borrow heavily from the techniques of linear programming, widely used in industry, business, and commerce to optimize processes and products. These techniques have been around for a long time, and to implement them, we can resort to commercial computer software not yet discovered by the testing industry. In a sense, this book does not offer anything new. Then, to demonstrate the techniques's applicability, we had to reconceptualize the process of test design, introduce a new language to deal with it, integrate the treatment of content and statistical requirements for tests, and formulate typical test-design goals and requirements as simple linear models. More importantly, we also had to demonstrate the power and nearly universal applicability of these models through a wide range of empirical examples dealing with several test-design problems.

Although the topic of this book is *test design*, the term is somewhat ambiguous. The only stage in the design process at which something is actually designed is the second stage, item-pool design. From that point on, the production of a test only involves its assembly to certain specifications from a given item pool. The stages of item-pool design and test assembly can be based on the same techniques from linear programming. But these techniques are much more easily understood as tools of test assembly, and for didactic reasons, I first treat the problem of test assembly and return to the problem of item-pool design as one of the last topics in this book.

In particular, the book is organized as follows. Chapter 1 introduces the current practice of test development and explains some elementary concepts from test theory, such as reliability and validity, and item and test information. Chapter 2 introduces a standard language for formulating test specifications. In Chapter 3, I show how this language can be used to model test assembly problems as simple linear models. Chapter 4 discusses general approaches available in mathematical programming, more specifically integer or combinatorial programming, to solve these models. A variety of empirical examples of the applications of the techniques to test-assembly problems, including such problems as IRT-based and classical test assembly, assembling multiple test forms, assembling tests with item sets, multidimensional test assembly, and adaptive test assembly, are presented in Chapters 5–9. The topic of item-pool design for programs with fixed and adaptive tests is treated in Chapter 10 and 11, respectively. The book concludes with a few more reflective observations on the topic of test design.

My goal has been to write a book that will become a helpful resource on the desk of any test specialist. Therefore, I have done my utmost to keep the level of technical sophistication in this book at a minimum. Instead, I emphasize such aspects as problem analysis, nature of assumptions, and applicability of results. In principle, the mathematical knowledge required to understand this book comprises linear equalities and inequalities from high-school algebra and a familiarity with set theory notation. The few formulas from test theory used in this book are discussed in Chapter 1. In addition, a few concepts from linear programming that are required to understand our modeling approaches are reviewed in Appendix 1. Nevertheless, Chapter 4 had to be somewhat more technical because it deals with methods for solving optimization problems. Readers with no previous experience with this material may find the brief introductions to the various algorithms and heuristics in this chapter abstract. If they have no affinity for the subject, they should read this chapter only cursorily, skipping the details they do not understand. They can do so without losing anything needed to understand the rest of the book. Also, it is my experience that the subject of multidimensional test assembly in Chapter 8 and, for that matter, the extension of adaptive test assembly to a multidimensional item pool in the last sections of Chapter 9, is more difficult to understand, mainly because the generalization of the notion of information in a unidimensional test to the case of multidimensionality is not entirely intuitive. Readers with no interest in this subject can skip this portion of the book and go directly to Chapter 10, where we begin our treatment of the subject of item-pool design.

Although this book presents principles and techniques that can be used in the three stages of test specification, item-pool design, and test assembly, the stage of item-pool development is hardly touched. The steps of item pretesting and calibration executed in this stage are treated well in several other books and papers (e.g., Hambleton & Swaminathan, 1985; Lord, 1980; Lord & Novick, 1968), and it is not necessary to repeat this material here. As for the preceding step of writing items for a pool, I do go as far as to show how blueprints for items can be calculated at the level of specific item writers and offer suggestions on how to manage the item-writing process (Chapter 10). But I do not deal with the actual process of item writing. Current item-writing practices are challenged by rapid developments in techniques for algorithmic item writing (e.g., Irvine & Kyllonen, 2002). I find these developments, which are in the same spirit as the "engineering approach" to test design advocated in this book, most promising, and I hope that, before too long, the two technologies will meet and integrate. This integration would reserve the intellectually more challenging parts of test design for our test specialists and allow them to assign their more boring daily operations to computer algorithms.

Several of the themes in this book were addressed in earlier research projects at the Department of Research Methodology, Measurement, and Data Analysis at the University of Twente. Over a period of more than 15 years, I have had the privilege of supervising dissertations on problems in test assembly and item-pool design by Jos J. Adema, Ellen Timminga, Bernard P. Veldkamp, and, currently, Adelaide Ariel. Their cooperation, creativity, and technical skills have been greatly appreciated. Special mention is deserved by Wim M.M. Tielen, who as a software specialist has provided continuous support in numerous test-assembly projects.

The majority of the research projects in this book were done with financial support from the Law School Admissions Council (LSAC), Newtown, Pennsylvania. Its continuous belief in what I have been doing has been an important stimulus to me, for which I am much indebted to Peter J. Pashley, Lynda M. Reese, Stephen T. Schreiber, and Philip D. Shelton. My main contact with the test specialists at the LSAC was Stephen E. Luebke, who provided all of the information about the item pools and test specifications that I needed for the projects in this book.

This book was written while I was a Fellow of the Center for Advanced Study in the Behavioral Sciences, Stanford, California. My fellowship was supported by a grant to the Center from the Spencer Foundation, for which I am most grateful. The tranquil location of the Center, on the top of a hill just above the Stanford campus, and the possession of a study overlooking a beautiful portion of the Santa Cruz Mountains, enabled me to view things in a wide perspective. I thank Doug McAdam, Director, and Mark Turner, Associate Director, as well as their entire staff, for their outstanding support during my fellowship. I am indebted to Kathleen Much for her

editorial comments on a portion of this book as well as on several other papers I wrote while at the Center.

Seven chapters of this book were tried out in a course on advanced topics in educational measurement at Michigan State University by Mark D. Reckase. His critical comments and those of his students led to many improvements in the original text. Bernard P. Veldkamp read several earlier versions of the manuscript and checked all exercises, while Adelaide Ariel went far beyond her call of duty with her help with the preparation of the graphs in this book. I am also grateful to Krista Breithaupt, Simon Bussman, Britta Colver, Alexander Freund, Heiko Grossman, Donovan Hare, Heinz Holling and Tobias Kuhn, whose comments helped me tremendously to polish the final version of the manuscript. The last chapter was completed while I enjoyed a fellowship from the Invitational Fellowship Program for Research in Japan at the University of Tokyo. I am indebted to the Japan Society for the Promotion of Science (JSPS) for the fellowship and to Kazuo Shigemasu for having been such a charming host.

Last but not least, I would like to thank John Kimmel, Executive Editor, Statistics, at Springer for being a quick and helpful source of information during the production of this book.

Each of the people whose support I acknowledge here have made my task as an author much more pleasant than I anticipated when I began working on the book.

Wim J. van der Linden
University of Twente

Acknowledgment of Copyrights

Several of the figures and tables in this book are (slightly re-edited) versions of figures and tables in earlier journal articles by the author. He is grateful to *Applied Psychological Measurement* for the right to reproduce Figures 5.6, 7.1, and 7.2 and Tables 5.3, 6.2, 6.3, 7.2, 10.1, and 11.1, to the *Journal of Educational and Behavioral Statistics* for the right to reproduce Figures 9.1, 11.1, and 11.2, and to the *Journal of Educational Measurement* for the right to reproduce Figures 11.3 and 11.4 and Table 5.1.

http://www.springer.com/978-0-387-20272-3

Linear Models for Optimal Test Design
van der Linden, W.J.
2005, XXIV, 408 p. 44 illus., Hardcover
ISBN: 978-0-387-20272-3