Statistics for Social and Behavioral Sciences

Advisors:
S.E. Fienberg W.J. van der Linden

Wim J. van der Linden

Linear Models for Optimal Test Design

Foreword by Ronald K. Hambleton

With 44 Figures

Springer

Wim J. van der Linden
Department of Measurement
and Data Analysis
Faculty of Behavioral Sciences
University of Twente
7500 AE Enschede
The Netherlands
w.j.vanderlinden@utwente.nl

Advisors:

Stephen E. Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA

Wim J. van der Linden
Department of Measurement
and Data Analysis
Faculty of Behavioral Sciences
University of Twente
7500 AE Enschede
The Netherlands

Library of Congress Control Number: 2005923810

ISBN-10: 0-387-20272-2
ISBN-13: 978-0387-20272-3

Printed on acid-free paper.

Printed in the United States of America. (EB)

9 8 7 6 5 4 3 2 1

springeronline.com

Voor mijn lieve Tonneke

Foreword

Over my nearly forty years of teaching and conducting research in the field of psychometric methods, I have seen a number of major technical advances that respond to pressing educational and psychological measurement problems. The development of criterion-referenced assessment was the first, beginning in the late 1960s with the important work of Robert Glaser and Jim Popham, in response to the need for assessments that considered candidate performance in relation to a well-defined body of knowledge and skills rather than in relation to a norm group. The development of criterion-referenced testing methodology with a focus on decision-theoretic concepts and methods, content validity, standard-setting, and the recognition of the merits of both criterion-norm-referenced and criterion-referenced assessments has tremendously influenced current test theory and testing .

The second major advance was the introduction of item response-theory (IRT) and associated models and their applications to replace classical test theory (CTT) and related practices. Beginning slowly in the 1940s and 1950s with the pioneering work of Frederic Lord, Allan Birnbaum, and Georg Rasch, by the 1970s the measurement journals were full of important research studies describing new IRT models, technical advances in model parameter estimation and model fit, and research on applications of IRT models to equating, test development, the detection of potentially biased test items, and adaptive testing. The overall goal has been to improve and expand measurement practices by overcoming several shortcomings of classical test theory: dependence of test-item statistics and reliability estimates on examinee samples, dependence of examinee true score estimates on the particular choices of test items, and the limitation in CTT of modeling ex-

aminee performance at the test level rather than at the item level. The last two shortcomings are especially problematic for adaptive testing, where it is important to be able to assess ability independently of particular test items and closely link item statistics to examinee ability or proficiency for the optimal selection of test items to shorten testing time and improve measurement precision on a per item basis. Today, the teaching of item-response theory is common in graduate training programs in psychometric methods, and IRT models and applications dominate the field of assessment.

The third major advance was the transition of testing practices from the administration of tests via paper and pencil to administration via the computer. This transition, which began in the late 1970s in the United States with considerable research funding from the armed services and with the leadership of such important scholars as Frederic Lord, Mark Reckase, Howard Wainer, and David Weiss, is widespread, with hundreds of credentialing exams (e.g., the Uniform Certified Public Accountancy Exams, the nursing exams, and securities industry exams in the United States), admissions tests (e.g., the Graduate Record Exam, the Graduate Management Admissions Test, and the Test of English as a Foreign Language), and achievement tests (e.g., high-school graduation tests in Virginia) being administered to candidates via computers, with more tests being added every month. The computer has added flexibility (with many testing programs, candidates can now take tests when they feel they are ready or when they need to take the tests), immediate scoring capabilities (thus removing what can often be months of waiting time for candidates), and the capability of assessing knowledge and skills that could not be easily assessed with paper-and-pencil tests. On this latter point, higher-level thinking skills, complex problem-solving, conducting research using reference materials, and much more are now being included in assessments because of the power of the computer.

Assessing candidates at a computer is becoming routine, and now a number of very important lines of research have been initiated. Research on automated scoring of constructed responses will ensure that computer-based testing can include the free-response test-item format, and thus the construct validity of many assessments will be enhanced. Research on automated item generation represents the next stage in test-item development and should expedite item writing, expand item pools, and lower the costs of item development. Automated item generation also responds to one of the main threats to the validity of computer-based testing with flexible candidate scheduling, and that is the overexposure of test items. With more test items available, the problem of overexposure of test items will be reduced.

Perhaps the most researched aspect of computer-based testing concerns the choice of test design. Initially, the focus was on fully adaptive tests. How should the first test item be selected? How should the second and third items and so on, be selected? When should testing be discontinued? How should ability or proficiency following the administration of each item be

estimated? Other test designs have been studied, too: multistage computer-based test designs (instead of selecting one optimal item after another, a block of test items, sometimes called "testlets" or "modules" are selected in some optimal fashion), and linear on-the-fly test designs (random or adaptive selection of tests subject to a variety of content and statistical constraints). Even the conventional linear test has been popular with one of a number of parallel forms being selected at random for administration to a candidate at a computer. But when computer-based testing research was initiated in the late 1970s, aptitude testing was the focus (e.g., the Armed Services Vocational Aptitude Battery), and detailed content-validity considerations were not a central concern. As the focus shifted to the study of computer-based achievement tests and credentialing exams (i.e., criterion-referenced tests) and the use of test scores became more important (e.g., credentialing exams are used to determine who is qualified to obtain a license or certificate to practice in a profession), content considerations became absolutely central to test defensibility and validity, and balancing tests from one examinee to the next for the length of item stems, the balance of constructed and selected response items, minimizing the overuse of test items, meeting detailed content specifications, building tests to match target information functions, and more, considerably more sophisticated methods for item selection were needed. It was in this computer-based testing environment that automated test assembly was born.

I have probably known about automated test assembly since 1983 (Wendy Yen wrote about it in one of her many papers), but the first paper I recall reading that was dedicated to the topic, and it is a classic in the psychometric methods field today, was the paper by Professor Wim van der Linden and Ellen Boekkooi-Timminga published in *Psychometrika* in 1989. In this paper, the authors introduced the concepts underlying automated test assembly and provided some very useful examples. I was fascinated that just about any content and statistical criteria that a test developer might want to impose on a test could be specified by them in the form of linear (in)equalities. Also, a test developer could choose an "objective function" to serve as the goal for test development. With a goal for test development reflected in an "objective function," such as with respect to a target test-information function (and perhaps even several goals), and both content and statistical specifications described in the form of linear constraints, the computer could find a set of test items that maximally met the needs of the test developer. What a breakthrough! I might add that initially there was concern by some test developers that they might be losing control of their tests, but later it became clear that the computer could be used to produce, when desired, first drafts of tests that could then be reviewed and revised by committees.

The 1989 van der Linden and Boekkooi-Timminga paper was the first that I recall that brought together three immensely important technologies, two that I have already highlighted as major advances in the psychometric

methods field—item-response theory and the use of the computer—and also operations research. But what impresses me today is that automated test assembly impacts or capitalizes on all of the major advances in the last 40 years of my career: criterion-referenced and norm-referenced assessments, item-response theory, computer-based testing, and new computer-based test designs, as well as emerging new assessment formats.

By 2004, I had accumulated a hundred papers (and probably more) on the topic. Most are by Professor Wim van der Linden and his colleagues in the Netherlands, but many other researchers have joined in and are producing important work and advancing the field. These papers overflow my files on item-response theory, test design, computerized adaptive testing, item selection, item-bank inventory, item-exposure controls, and many more topics. My filing system today is simply not capable of organizing and sequencing all of the contributions on the topic of automated test assembly since 1989, and I have lost track of the many lines of research, the most important advances, and so on. Perhaps if I were closely working in the field, the lines of research would be clearer to me, but like many measurement specialists, I have a number of research interests, and it is not possible today to be fully conversant with all of them. But from a distance, it was clear to me that automated test assembly, or optimal test design, or automated test construction, all terms that I have seen used in the field, was going to provide the next generation of test-design methods—interestingly whether or not a test was actually going to be administered at a computer! Now, with one book, van der Linden's *Linear Models for Optimal Test Design*, order in my world has been restored with respect to this immensely important topic, and future generations of assessment specialists and researchers will benefit from Professor Wim van der Linden's technical advances and succinct writing skills.

I believe *Linear Models for Optimal Test Design* should be required reading for anyone seriously interested in the psychometric methods field. Computers have brought about major changes in the way we think about tests, construct tests, administer tests, and report scores. Professor van der Linden has written a book that organizes, clarifies, and expands what is known about test design for the next generation of tests, and test design is the base or centerpiece for all future testing. He has done a superb job of organizing and synthesizing the topic of automated test assembly for readers, providing a step-by-step introduction to the topic, and offering lots of examples to support the relevant theory and practices. The field is much richer for Professor van der Linden's contribution, and I expect this book will both improve the practice of test development in the future and spur others to carry out additional research.

Ronald K. Hambleton
University of Massachusetts at Amherst

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.

Contents

1

Brief History of Test Theory and Design

Standardized testing was common practice in some ancient cultures long before western civilization developed—a well-known example is nationwide testing for civil service in ancient China. But we had to wait until the early twentieth century before it was introduced in western psychology. In 1905, Binet and Simon developed their intelligence test to identify students with mental retardation in Paris schools (Binet & Simon, 1905). Remarkably, this test already had most of the features characteristic of modern adaptive testing. The test was meant for individualized administration with a human proctor who scored the students during the test and selected the items. Standardization was obtained through the use of the same item pool and the application of the same detailed rules of item selection and scoring for all test takers.

The idea of standardized testing was extended from individualized testing to group-based, paper-and-pencil testing later in the twentieth century. The main stimuli for this transition were the necessities of placing large numbers of conscripts in the U.S. army during World Wars I and II and of fair admission methods to regulate the huge increase in student inflow into higher education in the second half of the twentieth century. These developments led to the large-scale use of multiple-choice tests—the ultimate format with objective, machine-based scoring of the test takers' responses to the test items.

In the early 1970s, a different type of testing emerged, first exclusively in education but later also in psychology. This new development was motivated by attempts to improve student learning in schools through frequent feedback on their achievements by tests embedded in the instruction. The

first idea was to offer students self-paced routes through series of small instructional modules, each finishing with a mastery test. Later, this idea was extended with choices between alternative modes of learning and students working more freely on series of assignments. A natural consequence of this development for individualized instruction was the need for item banking to support testing on demand (also referred to as "walk-in testing"). As a result, the earlier notion of a standardized test as the same paper-and-pencil form for each test taker evolved into the idea of testing from item pools defined by extensive lists of specifications and algorithmic item writing and test assembly. The advent of cheap personal computers with plentiful computational power in the early 1990s stimulated these changes enormously. When a few years later the technology of item banking and individualized testing matured and eventually led to the large-scale introduction of computerized adaptive testing in education, it began to find applications in psychological testing as well.

It is remarkable how these developments have their parallels in two different periods in the history of testing. The first period covers the first half of the twentieth century, when classical test theory (CTT) was developed. This theory mainly supports standardized testing with a group-based paper-and-pencil test for a fixed population of test takers. In the 1950s, ideas for a new test theory were explored and a second period began, in which item-response theory (IRT) was developed. It received its first comprehensive formalization in the late 1960s, a more thorough statistical treatment in the 1970s–1980s, and began to be applied widely in the 1990s. As a matter of fact, it is still in the process of being extended, particularly into the direction of models for more complicated response formats, models with more comprehensive parameterization (for instance, to deal with background variables of the test takers, sophisticated sampling designs, and multidimensional abilities), and models for response times. The introduction of IRT has been critical to the development of the new technology of item banking and individualized testing. Also, IRT allows for item formats that are closer to the current instructional requirements and relies heavily on the (real-time) use of the computational power provided by modern computers.

In the next sections of this chapter, we review these two stages in somewhat more detail and introduce the basic concepts in test development and test theory on which this book relies.

1.1 Classical Test Design

1.1.1 Standardized Testing in Psychology

Classical test design has been strongly dominated by the idea of a standardized test developed in psychology. Psychological tests are typically

produced as an *encore* to a development in psychological theory. The result of such a development is a theoretical network around one or more new constructs, for example, certain special abilities, personality traits, or psychodiagnostic dimensions. Test development begins if more systematic empirical research is needed to test hypotheses on these constructs against empirical reality.

As a result, psychological tests are seldom developed by test specialists but mostly by psychologists familiar with the research on the constructs for which they are to be used as a measurement instrument. These researchers use their knowledge to design the tasks or items in the test and to choose the rules for scoring them. Usually, the items are written together as a set that is assumed to cover the construct best. Typically this set is somewhat larger than actually needed, to allow for a possible failure of some of the items during pretesting.

This developmental process can be characterized as a one-shot approach based on the best theories and insights available at the time. New items are written and tried out only if a new version of the test has to be developed, which happens if new insights and progress in psychological theory make the current version obsolete. The same psychological test can be easily used for over a decade before the need for a subsequent version is felt.

Empirical pretesting of items usually serves a threefold purpose. First, it allows for a screening of estimates of the item parameters and the possible removal of items with estimates suggesting undesirable behavior. The parameters used in a classical item analysis are briefly reviewed in the next section. Second, predictions following from the theory underlying the constructs are confronted with empirical data. These predictions may be on the correlational structure of the test scores with other measures in the study (for example, in a multitrait-multimethod study) or on differences between the score distributions of certain groups of persons. The results from this part of the study are used both to test the psychological theory and validate the test. Third, the test is normed for its intended population of persons. This part of the tryout involves extensive sampling of the population and the estimation of a norm table for it. If a new version of an existing test is pretested, the data are used for score equating. The goal then is to estimate the transformation that maps the score scale of the new version of the test to the scale of the old version. This transformation generates the same norm table for both versions. To the knowledge of the author, the first large-scale study with this type of score equating ever was for the new version of *Wechsler-Bellevue Intelligence Scale* in 1939.

This process of development of a standardized test has a more than superficial relation with CTT. In the next section, we review a few basic concepts from CTT. These concepts will be used later in this book and will also help us to discuss the close relation between test theory and design in a subsequent section.

1.1.2 Classical Test Theory

The core of classical test theory (CTT) is a two-level model that decomposes the observed test scores into so-called true scores and errors. The presence of two levels in the model is due to the fact that CTT addresses both the case of a fixed test taker and a random person sampled from a population. At either level, the test is considered as fixed; for instance, the case of testing with random samples of items is not addressed.

Fixed Person

Let X_{jt} be the observed score of fixed person j on test t. A basic assumption in CTT is that this observed score, which can be any quantity defined on the item scores of the person, is a *random variable*. The assumption reflects the belief that if we replicated the test several times, a distribution of outcomes would be observed. This experiment can actually be done for tests of stable physical abilities, for which memory and learning do not play a role, but is hypothetical for the more mental and cognitive abilities. Although X_{jt} is random, the shape of its distribution is unknown. In fact, the goal of test theory is to provide models that help us make inferences of the properties of this distribution from actual observed scores of the person.

Observed score X_{jt} can be used to define two new quantities:

$$\tau_{jt} = \mathcal{E}X_{jt}, \tag{1.1}$$

$$E_{jt} = X_{jt} - \tau_{jt}. \tag{1.2}$$

The first quantity is the *true score* for person j on test t, τ_{jt}, which is defined as the expected value or mean of the observed-score distribution. The second is the *error* in the observed score, E_{jt}, which is defined as the difference between the person's observed score and true score. Both definitions are motivated by practical considerations only; if we have to summarize the distribution of the observed score by a single fixed parameter, and the distribution is not known to be skewed, it makes sense to choose its mean, and if an actual observation of X_{jt} is used to estimate this mean, we make an error equal to E_{jt}.

The definitions in (1.1) and (1.2) imply the following model for the score of a fixed person:

$$X_{jt} = \tau_{jt} + E_{jt}. \tag{1.3}$$

This model is nothing but a convenient summary of the preceding introduction. The only assumption underlying it is the randomness of the observed score X_{jt}; the fact that the true score and error are combined additively does not involve anything new above or beyond the definition of these two quantities.

Random Person

If the persons are sampled randomly from a population, the true score also becomes random. In addition, the observed score and error contain two levels of randomness, one level because we sample a person from the population and another because we sample an observed score from the person's distribution. Let J represent the random person sampled from the population and T_{Jt} the random true score. The model in (1.3) becomes:

$$X_{Jt} = T_{Jt} + E_{Jt}. \tag{1.4}$$

Again, the only new assumption underlying this extension of the model is on the random status of a variable—this time the true score; no assumption of linearity whatsoever has been made.

Item and Test Parameters

One of the major roles of CTT is as a producer of meaningful parameters for item and test analysis. All parameters reviewed in this section are at the level of the population model in (1.4).

A key parameter is the *reliability coefficient* of the observed score X_{Jt}, usually (but incorrectly) referred to as the reliability of the test instead of a score. This parameter is defined as the squared (linear) correlation coefficient between the observed and true scores on the test, ρ^2_{TX}. (Because the level of modeling is now well understood, we henceforth omit the indices of the scores where possible.)

The choice of the correlation between X and T is intuitively clear: If $X = T$ for the population of persons, (1.4) shows that X does not contain any error for each of them, and the correlation between X and T is equal to 1. Likewise, it is easy to show that if $X = E$ (that is, X contains only error for each person), the correlation is equal to 0.

The fact that we do not define reliability as the correlation coefficient between X and T but as the square of it is to treat ourselves to another useful interpretation. A standard interpretation of a squared correlation coefficient is as a proportion of the explained variance. Analogously, in CTT, the reliability coefficient can be shown to be equal to

$$\rho^2_{TX} = \frac{\mathrm{Var}(T_{Jt})}{\mathrm{Var}(X_{Jt})}, \tag{1.5}$$

which is the proportion of of the true-score variance relative to the observed-score variance in the population of persons. This equality thus shows that the true-score variance in a population of persons can be conceived of as the proportion of observed-score variance explained by the differences in true scores between the persons.

If test scores are used to predict a future variable, Y (for example, success in a therapy or training program), the reliability coefficient remains a key

parameter, but the correlation of observed score X with Y, instead of with its true score T, becomes the ultimate criterion of success for the test. For this reason, we define the *validity coefficient* of a test score X as its (linear) correlation with criterion Y, ρ_{XY}.

Observe that, unlike the reliability coefficient, the validity coefficient is not a squared correlation coefficient. The reason for this lies in the following two results for the reliability coefficient that can be derived from the model in (1.4). First, using well-known rules for variances and covariances, it can be shown that if X and X' are the observed scores on two replications of the test for the same persons, it holds that

$$\rho_{XT}^2 = \rho_{XX'}. \tag{1.6}$$

This result is most remarkable in that it shows that the reliability coefficient, which is the squared correlation between the observed scores and their unobservable true scores, is equal to the correlation between two replications of the observed scores. Likewise, it can be shown that

$$\rho_{XT} \geq \rho_{XY} \tag{1.7}$$

for any score Y. The result (1.7) tells us that the predictive validity coefficient of a test can never exceed the correlation between its observed score and true scores; or, the other way around, the observed score on a test is always the best "predictor" of its true score. Observe that (1.7) also relates the correlation of an unobservable score to the correlation between two observed scores.

An important item parameter in CTT is the *item difficulty* or π *value.* Let U_i be the score on item i in the test, with $U_i = 1$ the value for a correct response and $U_i = 0$ the value for an incorrect response. The classical difficulty parameter of item i is defined as the expected value or mean of U_i in the population of persons

$$\pi_i = \mathcal{E}U_i. \tag{1.8}$$

CTT also has an *item-discrimination parameter,* which is defined as the correlation between the item score and the observed test score

$$\rho_{iX} = \mathrm{Cor}(U_i, X) = \frac{\sigma_{iX}}{\sigma_i \sigma_X}, \tag{1.9}$$

where σ_{iX}, σ_i, and σ_X are the covariance between U_i and X, and the standard deviations of U_i and X, respectively. Obviously, a large value for ρ_{iX} implies a score on item i that discriminates well between persons with a high and a low total score on the test; hence the name "discrimination parameter." Recall, however, that X is composed of the scores on all items in the test; it is therefore somewhat misleading to view a correlation between U_i and X as an exclusive property of item i.

Analogously to (1.9), we define the correlation between the score on item i and the observed score Y on a success criterion,

$$\rho_{iY} = \mathrm{Cor}(U_i, Y) = \frac{\sigma_{iY}}{\sigma_i \sigma_Y}, \tag{1.10}$$

as the *item validity* or the item-criterion correlation for item i. It represents how well score U_i discriminates between persons with high and low scores on criterion Y in a predictive validity study.

All the parameters above were defined as population quantities. They can be estimated directly by their sample equivalents, with the exception of the reliability coefficient, which is based on the correlation with an unobservable true score. The equality in (1.6) suggests estimating the reliability coefficient by the sample correlation between observed scores X and X' on two replicated administrations of the test. But, in practice, due to learning and memory effects, it is seldom possible to realize two exact replications.

An alternative is to use the inequality

$$\rho_{XT}^2 \geq \alpha, \tag{1.11}$$

which can be derived from the model in (1.4), where *coefficient* α is defined as

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_X^2}\right] \tag{1.12}$$

and n is the length of the test. Coefficient α is a coefficient for the internal consistency of a test; that is, the degree to which all item scores in a test correlate positively with one another. The relation in (1.11) thus shows that the reliability of an observed score can never be smaller than the internal consistency of the item scores on which it is calculated. Coefficient α can be estimated in a single administration of the test; it only contains the item variances, σ_i^2, and the total observed-score variance, σ_X^2, which can be estimated directly by their sample equivalents. If the test approximates the ideal of a unidimensional test, the error involved in the estimation of ρ_{XT}^2 through α tends to be small.

It is helpful to know that the following relation holds for the standard deviation of observed score X:

$$\sigma_X = \sum_{i=1}^{n} \sigma_i \rho_{iX}. \tag{1.13}$$

Replacing σ_X^2 in (1.12) by the square of this sum of products of item parameters leads to:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\left(\sum_{i=1}^{n}\sigma_i\rho_{iX}\right)^2}\right] \tag{1.14}$$

Except for the (known) test length n, this expression for α is entirely based on two item parameters, σ_i and ρ_{iX}. It allows us to calculate how the removal or addition of an item to the test changes the value of α.

For the validity coefficient, we are also able to derive an expression based entirely on sums of item parameters. The expression is

$$\rho_{XY} = \frac{\sum_{i=1}^{n}\sigma_i\rho_{iY}}{\sum_{i=1}^{n}\sigma_i\rho_{iX}}. \tag{1.15}$$

It shows us how the predictive validity of a test is composed of the item variances, item-discrimination parameters, and item validities.

We will rely heavily on the expressions in (1.14) and (1.15) when we discuss models for classical test assembly in Section 5.2.

1.1.3 Discussion

Classical test design and classical test theory are different sides of the same coin. Both are based on identical methodological ideas, of which the notion of standardization is the core.

When a test is designed, the conditions in the testing procedure that determine the ability to be tested are standardized. Standardization implies the same conditions over replications. The definition of the observed score in CTT as random over replications of the test is entirely consistent with this idea of standardization. The mean of the distribution of this score is a fixed parameter that summarizes the effects of all standardized conditions. It seems natural to call this mean the true score. The error score summarizes the effects of conditions that have been left free. Because these effects are random across replications, the error score is random.

At approximately the same time as the introduction of classical test theory, similar notions were developed in the methodology of experimental design, with its emphasis on manipulation and randomization. In fact, just as CTT is the statistical analog of standardized testing, the analog of experimental design is analysis of variance. It is therefore not surprising that strong parallels exist between the linear models in (1.3) and (1.4) and some models in analysis of variance.

The assumption of sampling from a fixed population is another common characteristic of classical test design and CTT. For example, one of the main goals of psychological testing is to estimate the test taker's relative standing in this population, often with the intention of seeing if this person belongs to the "normal" portion of the population distribution or an "abnormal" tail. The interest in norm tables is a logical consequence of this goal. In CTT, this interest finds its parallel in the assumption of random sampling of persons from a fixed population.

To get an accurate estimate of the true scores of a population of test takers, the test should be designed to discriminate maximally between as many persons in the population as possible. Statistically, this goal is realized best by a test with its π values close to .50 and values for the item-discrimination parameter ρ_{iX} as large as possible. This choice of parameter values has been the standard of the testing industry for a long time. The fact that these parameters can be interpreted only for a population of persons was not observed to be a hindrance but was a prerequisite according to the prevalent conception of testing (Exercise 1.1).

The classical conception of test development involved no stimulus to item banking whatsoever. If the test items are the best given the current state of psychological theory and have been shown to meet the statistical requirements for the intended population, there is no need whatsoever to write more items. Producing more can only lead to an increase in quality. The only reason to write new items is if the test becomes obsolete due to new developments in psychological theory.

It is not our intention to suggest that this classical complex is wrong. On the contrary, it is coherent, well-developed, and statistically correct. If a single test for a fixed population has to be developed, and the interest is exclusively in estimating score differences in a population of persons, the combination of classical test design and classical test theory is still a powerful choice. The methodology offered in this book can also be applied to classical test design (see Section 5.2).

But if testing has to serve a different goal, another choice of test-design principles and theory has to be made. As discussed in the next section, this was precisely what happened when testing was applied to instructional purposes.

1.2 Modern Test Design

1.2.1 New Notion of Standardization

The first large-scale use of educational tests was for admission to higher education. For this application, the assumption of a fixed population still made sense, but the assumption of a fixed test involved going through the whole cycle of test development on an annual basis. This requirement put

a serious claim on the resources of the testing organizations. They soon discovered that it was more efficient to use item banking. In *item banking,* test items are written and pretested on a more continuous basis, and tests are assembled from the pool of items currently present in the item-banking system.

The need for a new test theory was felt more seriously when the use of tests for instructional purposes was explored, particularly when the ideas moved into the direction of individualized instruction. The assumption that students are sampled from a fixed population does not make much sense if individual students take different instructional routes. In fact, it is even inconsistent with the notion of learning at all. A score distribution of a population of students can only remain fixed if their abilities are—not if they develop as a result of learning and forgetting. Likewise, the idea of a single best test soon had to be killed. If students are tested individually and at different levels of development, larger numbers of tests with measurement qualities geared to the individual student's level are necessary.

If the assumptions of a fixed population and a single best test have to be dropped, other features of the classical complex become problematic, too. For example, classical item and test parameters, such as the π value, item-discrimination parameter, and reliability coefficient, are based on the assumption of a fixed population and lose their meaning if no such population exists. Likewise, the definition of the true score in CTT is based on the assumption of a single fixed test. If different students take different tests, their number-correct scores are no longer comparable. Also, if the same student is retested using different tests, it is impossible to use this score for monitoring what this person has learned.

It is obvious that with the emergence of these newer types of testing, a new test theory was required. Item-response theory (IRT), of which the key concepts are introduced in the next section, has filled the void. It is not for a fixed test for a fixed population but for a pool of items measuring the same ability and for individual persons demonstrating the ability in their responses to these items. It also offers us the tools for calibrating items taken by different persons on a fixed scale. In addition, item parameters in IRT describe the properties of the items relative to this scale instead of a population of persons. Therefore, these parameters can be used to assemble a test that is locally best (i.e., has optimal accuracy at the person's ability level). They can also be used to score persons on the same scale, no matter what test they take from the pool.

In fact, the emergence of these newer types of testing and the simultaneous development of IRT have led to the replacement of the "classical complex" in testing in Section 1.1.3 by a new paradigm. The core of this paradigm is a changed notion of standardization. To standardize a test, it is no longer necessary to give each person an identical set of items (or, for that matter, test them under identical conditions). It is sufficient that the items be written to explicit content specifications and that the remaining differ-