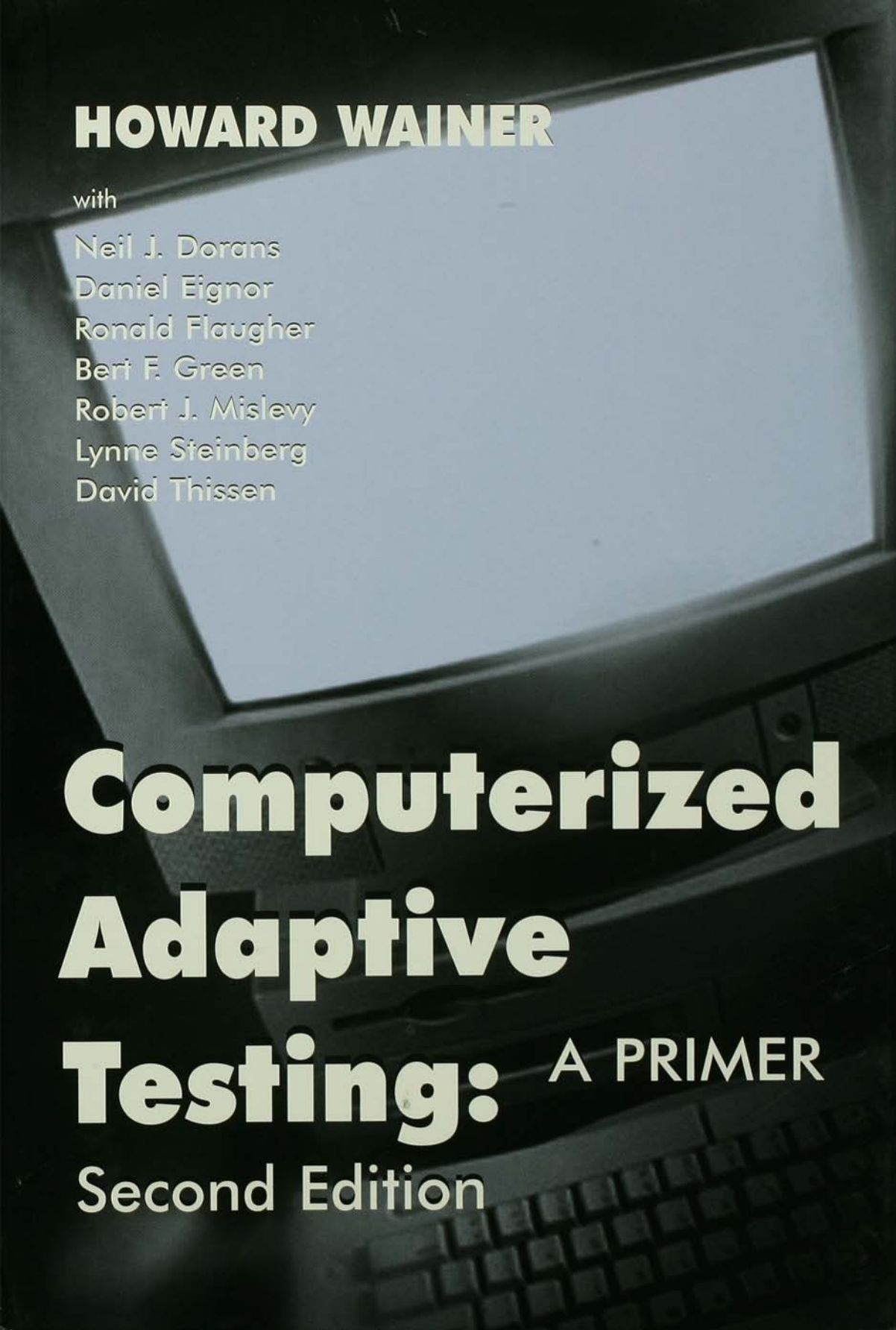

HOWARD WAINER

with

Neil J. Dorans
Daniel Eignor
Ronald Flaugher
Bert F. Green
Robert J. Mislevy
Lynne Steinberg
David Thissen

Computerized Adaptive Testing: A PRIMER

Second Edition

Computerized Adaptive Testing
A Primer
Second Edition

This page intentionally left blank

Computerized Adaptive Testing

A Primer

Second Edition

by
Howard Wainer
Educational Testing Service

with

Neil J. Dorans
Educational Testing Service

Daniel Eignor
Educational Testing Service

Ronald Flaugher
Educational Testing Service

Bert F. Green
Johns Hopkins University

Robert J. Mislevy
Educational Testing Service

Lynne Steinberg
Indiana University

David Thissen
University of Kansas

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2000 by Lawrence Erlbaum Associates, Publishers
Published 2014 by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN
711 Third Avenue, New York, NY 10017, USA
Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data

Wainer, Howard.
Computerized adaptive testing : a primer / by Howard Wainer with Neil J. Dorans . . . [et al.]. —2nd ed.
p. cm.
Includes bibliographical references (p.) and indexes.
ISBN 0-8058-3511-3 (cloth : alk. paper)
1. Computerized adaptive testing. I. Dorans, Neil J. II. Title.

LB3060.32.C65 W25 2000
371.26—dc21 99-087976

ISBN 978-0-805-83511-3 (hbk)

Contents

Foreword to the First Edition

C. Victor Bunderson
Vice President, Research Management
Educational Testing Service

Publication of this book signals the rite of passage of Computerized Adaptive Testing (CAT) from childhood into young adulthood. CAT is now a viable alternative form of measurement, backed by a rigorous technology involving far more than hardware and software.

The childhood of CAT was characterized by the publication of speculative articles about the future benefits of computerized testing, and this will no doubt continue. These articles wax enthusiastic about a technological future extrapolated from a scattered mosaic of technical articles and reports, not widely accessible, on different aspects of CAT. The mosaic of technical articles dealt with central issues concerning item response theory and its implementation on computers for adaptive testing.

This *CAT primer* puts the mosaic together. Although the book is technical when it must be, it is well written and interesting. A hypothetical test taken as a prelude to employment, the *Gedanken Computerized Adaptive Test* (GCAT) is used to tie the book together with a common example. It also highlights important features and problems presented by different types of computerized subtests. One GCAT subtest uses long paragraphs that might not fit on the computer screen, another involves computer presentation of graphics for measuring memory, and one deals with clerical response speed.

The lay reader will enjoy following the adventures of the bright and eager Cindy and her amiable friend, Scott—two young people who took the GCAT. Intrigued by her experience with the test, Cindy leads Scott (whose level of interest is delightfully higher than would be predicted by his GCAT score) and the reader from chapter to chapter to satisfy her seemingly insatiable interest in CAT arcana like equating and reliability.

It is not immediately clear why this book should be called a primer. It presents a rather thorough and up-to-date presentation of the state-of-the-art in the field. What is left unsaid has not yet been discovered, or at least has not been established rigorously.

It is an axiom in computer use that first we computerize what we know how to do without computers, then we discover applications unthought of before computerization. So it is with CAT. The GCAT presents a familiar purpose for testing: selection. It also assumes that there is a preexisting paper-and-pencil test battery and that the GCAT is a computerized version, offering the innovation of adaptive delivery of the sequence of items, immediate scoring, and adaptive determination of when to stop. Items, however, are still exclusively in the multiple-choice format, so important to cost-effective paper-and-pencil test scoring. The scaling methods are based on standard psychometrics, using an old metaphor for scaling: the meter stick (or 6 meter sticks, in the case of the GCAT). By assumption, these equal interval measuring sticks are used to measure relatively fixed and unchanging quantities of mental and psychomotor ability. These six latent traits are assumed to vary continuously in one dimension. But despite the close similarity to tests delivered by paper-and-pencil, the change to new possibilities predicted in the axiom of computer use is clearly underway.

One change is the move from items to testlets. Testlets, especially the branching kind, have not been a part of conventional testing. These have promise as replacements for individual items as the basic building blocks of CAT tests. Testlets serve as the means to create and control context effects and as a means to assure the fulfillment of content specifications. Response time measurement is also discussed, but not yet applied to the measurement of new constructs dealing with human processing speed.

Other changes should come as a result of wider use of the forms of computerized testing presented in this primer—such use would lead even further away from the forms, models, metaphors, and purposes of paper-and-pencil testing. A multitude of display and response formats and scaling metaphors are possible with tests administered by computer. Tests that present video motion and computer graphics and use these displays to depict process and change will offer new kinds of tasks—such tasks provide even more face validity and rich context than can be achieved with a testlet of multiple-choice items. Dynamic changes in displays introduce a new kind of adaptive test that permits changes in the features of a display based on the responses of the student—simulation tests.

The computer would also lead to response mode options well beyond the five-button key pad described in chapter 2. Students would be able to type in numbers, words, equations, sentences, and paragraphs. They would be able to enter responses by pointing a finger or mouse to parts of displays, or by marking or drawing. They would also be able to enter force and direction with joy sticks and other devices.

The equal interval meter stick for measuring fixed quantities of ability is not

the only metaphor for computerized measurement. This metaphor, and the assumptions of the IRT model, carry the implication of a stable and fixed quality to be measured. Other metaphors will be needed for dynamic measurement, where we assume the quantity to be measured is changing, as in learning.

One metaphor for dynamic measurement is a speedometer, a scale that shows a continuously varying quantity, used by the driver to control the speed of a vehicle. Another metaphor is a radar screen with icons identifying both the target and the attacking aircraft. This two-dimensional scaling of distance and direction can be used to gauge movement toward a goal. A user of this dynamic display can adjust the approaching object so that its trajectory converges on the target. The radar screen metaphor could have its counterpart in two-dimensional scales of use in education to monitor learning progress toward a goal. A new form of "cartography of intellectual territory" could utilize map displays on the computer screen with shaded or colored areas like those on charts representing isotherms, or those representing the distribution of commodities. The intellectual maps could display a representation of the existing distribution of human qualities over a domain, and could go a step further by depicting developing individual or group mastery of different topics or contents.

Another metaphor, sticking pins in a wall map, shows gradual coverage of territory as goals are accomplished in each part of the territory. Gradually filling in the domain map on the computer screen could give a learner or an operator a sense of progress toward completion.

Computerized measurement would thus evolve not only to new forms using advanced display and response functions, but to new purposes. These purposes will include continuous measurement of dynamically changing quantities in educational settings, and other purposes not possible with static paper-and-pencil delivery and static metaphors of measurement.

Current developers of interactive computer lessons and tests, excited by these novel possibilities for computerized testing and for testing integrated with training and instruction, have accelerated their creations into production without answering the questions of defensible measurement models, validity, reliability, and fairness discussed in this book. My breathless recital of new possibilities was not meant to put the field of computerized testing back into its childhood of speculative futures. It was meant to contrast the new applications even now developing with the standards in this book. Current practices frequently violate good measurement standards grossly (e.g., tests for selection or for high-stakes grading constructed by randomly pulling a fixed number of items from an item bank). My recital of coming developments was also meant as a call for standards of rigor and fairness in measurement and in use for these new kinds of tests and applications.

I hope the *CAT Primer* becomes a classic milestone in documenting a standard for solidly developing each subfield of computerized measurement. Other books about other forms of computerized testing will surely be written, and this fore-

word is an appeal that they meet standards not lower than those found in the *CAT Primer*. This book is based on 30 years of solid research and should, hopefully, influence future books in this field. It is sure to find its way into many classes in the field of educational measurement and onto many book shelves of both entering and practicing professionals in measurement. My best wish for it, however, is not that it remain a classic for 15 or 20 years, but that it be shown by the publication of other classics to have been a stimulus toward high standards of excellence and of equity in a broad and increasingly useful array of computerized measurement and instructional applications.

Foreword to the Second Edition

Drew H. Gitomer
Vice President, Research
Educational Testing Service

When the first edition of this volume was published, the authors cleverly illustrated their points with the fictional Gedanken Computerized Adaptive Test (GCAT). This year, ETS delivered its 1 millionth computer adaptive test. The promise of CAT, resulting from 30 years of foundational research, is now a cornerstone of standardized testing throughout the world.

And yet, we have only scratched the surface of the potential that the computer brings to assessment.

The earlier edition foresaw many of the issues that testing organizations would face as they implemented CAT on a broad scale. Complexities associated with test security, item pools, item selection, model fit, omitted responses, and timing constraints have all become quite evident during the last decade as large testing programs chose CAT as either an additional option or alternative to traditional paper-and-pencil testing. Indeed, the field is still grappling with these issues. The cautions expressed by the authors at the time were well placed and the broad implementation of CAT has allowed us to make significant progress in understanding and addressing a number of the issues they anticipated.

The new edition of this book is timely in that the authors (and consequently, the readers) have the benefit of considering the experience of operational CAT for several large admissions-type testing programs. The transition to CAT from paper-and-pencil testing has not been a seamless journey. This is not surprising given how CAT influences, and is influenced by, developments in technology, the economics of testing, and the perceptions and attitudes of test-takers. The authors do a thorough job of describing the principles and potential of CAT, as well as the potential pitfalls that those who enter the arena of CAT inevitably will confront.

The field of large-scale assessment has been conservative and this volume reflects that conservatism. By conservative, I do not mean to imply any political stance, but rather refer to practices that ensure that tests and test scores have consistent meanings over the years, that change comes slowly and with caution, and that incremental change is valued over any radical transformation. Thus, the authors focus primarily on current assessment practices, which to make the point, have not changed all that much during the decade between the first and second editions. They focus on issues associated with migrating existing paper-and-pencil instruments to CAT implementations. If done well, CAT can be more convenient for the test-taker in terms of scheduling, test-appropriateness, and score reporting. This volume provides excellent background in how to accomplish these goals as well as possible, from both a psychometric and operational perspective.

The conservatism of testing practice lies in sharp contrast with the context in which testing, and specifically CAT testing, takes place. Certainly, the Internet changes the equation dramatically. Access issues will rapidly become a non-issue, as the reliance on internal networks of testing machines is reduced. Certainly, however, security issues will continue to be at the forefront, though technology offers some intriguing new possibilities. Increasingly, familiarity with computers will also become a non-issue. Not only do most students have access to computers from an early age, but interfaces are becoming more sophisticated (i.e., transparent) as well, making it very easy for the unsophisticated user to accomplish tasks on a computer quite readily. And, of course, the economics of administering anything by computer are changing rapidly.

The purposes of testing are being expanded as well. There has been a loud and consistent call for testing to fulfill a more diagnostic and instructional role. Adaptive tests designed to understand what a student knows and does not know, and then provide instructional feedback require changes in how tests are developed, how student ability is modeled, how items are selected, and how information is reported. The authors endorse the use of CAT in these kinds of situations, especially when the stakes are such that security breaches are less likely. Throughout the volume, there are sections relevant to this evolution in testing, particularly discussions of multidimensional models of performance.

Tests designed to provide diagnostic feedback are likely to look very different from the tests that have been the focus of CAT to date. Tests are more apt to focus on more complex and integrated problems, with evidence of individual ability being inferred from information collected during problem solving. These kinds of complex assessments will challenge psychometric models that assume conditional independence. Such models are being developed by colleagues at ETS, and represent a new psychometrics that will undoubtedly play a role in the next generation of computerized adaptive testing. These tests will look far different from the tests that dominate the field today.

The needs of students are also changing. The demand for assessments associated with adult learning, training, and job qualification is expanding much more rapidly than the admissions testing market. CAT can play a significant role in providing useful and economically valuable information to test-takers and institutions alike. As the authors note, the relatively low volume of test takers sometimes makes the economics of developing such CAT tests uneconomical. However, as we make advances in our ability to automatically generate test items and as we increase our ability to automatically score complex responses, the feasibility of smaller volume CAT tests is likely to increase.

When we look at changes in assessment delivery brought about by the Internet, the focus on diagnostic and instructional assessments, as well as the rapid increase in nontraditional students and test use, it is clear that the constraints on CAT will change; some will be reduced and new ones will surface. None of these developments would be possible however, without the psychometric foundations of adaptive assessment that are described in this book. Furthermore, many of the challenges highlighted here will continue to confront the field, even as assessment practices are rethought.

I have no doubt that the second edition of this primer will continue to be an outstanding source of information about CAT, in the same ways that characterized the first edition. I see this volume as capturing the state of CAT as it exists today, yet intelligently conjecturing about issues facing CAT over the next decade. I also see the authors of this volume continuing to contribute to the psychometric evolution of CAT as technology and the purposes of testing continue to expand. If CAT is successful in achieving its promise and transforming educational assessment, then our assessments and the underlying psychometrics will have a very different look. This second edition of the CAT primer should keep us in good hands until the vision of a new generation of assessments is more fully realized.

This page intentionally left blank

Preface to the First Edition

This *Primer* came about because of a confluence of good fortune. First, because of a general interest at the Educational Testing Service in the use of the computer to improve the quality of testing. This resulted in a critical mass of researchers at ETS who had serious interest in computerized testing. Some of these researchers, with the full support of ETS management, prepared a proposal to the Naval Personnel Research and Development Center (NPRDC) in San Diego in response to their "Request For Proposals." This proposal offered to construct a *Technical Manual* in support of their Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) and to provide other services to the CAT-ASVAB program. Happily, our proposal was accepted and so began a long collaboration with NPRDC. The *Technical Manual* documented the work that had been done on CAT in general, and on the CAT-ASVAB in particular. It was (and is) a very long and technical document, which was (and is) incomplete. There is still much work remaining before the CAT-ASVAB is fully documented. Nevertheless, this technical manual provided for the first time a structured statement of what it takes to field a viable CAT. The CAT-ASVAB is a pioneering (even a visionary) effort, and therefore such a structured presentation provided a view of both what has been completed, and what remains yet to be done.

As the work on the manual was being completed, it seemed to me that with only a small amount of additional effort we could turn this very specific and technical document into a monograph of broad interest and usefulness. When I approached my colleagues (and coauthors) with the prospect of modifying their manual chapters into a *Primer* for CAT, I received a mixed reaction. The overall opinion was one of enthusiasm for the goals of such a project but disbelief about

the amount of effort I estimated that it would take. Their general estimate was an order of magnitude greater than what I had guessed at. As it turned out, this too was a healthy underestimate. However, the happy result of this extra effort has been a product of much higher quality than I had originally hoped for.

This *Primer* owes much to a variety of individuals. I would like to take this opportunity to try to thank each of them. The order of the thank-you's is, more-or-less, chronological.

First, my thanks to then Vice President of Research Management at ETS Ernest Anastasio, who was responsible for some of the enthusiasm among ETS researchers for computerized testing, and whose wisdom in allocating extra ETS resources aided in our winning the NPRDC contract.

Next to my ETS colleagues who aided in writing the original response to the NPRDC Request For Proposals. Two of these deserve special note. They are:

Martha Stocking, whose detailed and organized mind brought order to the proposal out of the chaos of our individual writings, and

Bill Ward, whose CAT experience and willingness to share it provided our proposal with a depth of knowledge that would have been difficult to obtain otherwise.

Next to the authors of the *CAT-ASVAB Technical Manual,* whose knowledge and hard work provided the grist from which this *Primer* was milled. They include:

Neil Dorans	Robert Mislevy
Benjamin Fairbank	Lynne Steinberg
Ronald Flaugher	Martha Stocking
David Hiester	David Thissen

Next to the NPRDC personnel, who did, sponsored, and/or supervised much of the research and work that we report here. Prominent among them are:

Bernard Rafacz, Gloria Jones-James and Elizabeth Wilbur for their major contributions to the design and development of the CAT-ASVAB microcomputer-based delivery system.

James McBride and Martin Wiskoff, who, during their tenure at NPRDC, gave the CAT-ASVAB Program its initial impetus, direction, and support.

Rebecca Hetter, Kathleen Moreno, Daniel Segall, J. Bradford Sympson and John Wolfe for their fundamental contributions to psychometric research that underlay much of the development and design of the CAT-ASVAB.

Drew Sands, Officer-in-Charge of the Joint Service CAT-ASVAB Program, and Director of the Testing Systems Department at NPRDC; Frank Vicino, Head of the Research Division in the department; and Jules Borack, Head

of the Systems Division. These individuals manage and supervise the research and development for the CAT-ASVAB Program at NPRDC.

Special appreciation and thanks go to Mary Schratz, the Contracting Officer's Technical Representative of the ETS support contract for NPRDC who oversaw all of the writing, provided much of the reference material, and read every word. Her comments made major improvements in the *Technical Manual.*

Next, I would like to thank the various advisory boards associated with the CAT-ASVAB. The membership of these panels has changed several times, and so naming everyone is beyond my ken. I would, however, like to single out Bruce Bloxom and Malcolm Ree for special thanks. Both of them spent many hours with me discussing adaptive testing. Their suggestions were always wise.

The writing of this *Primer* would have been impossible without corporate support from ETS. Absolutely indispensable in obtaining this support is ETS's current Vice President for Research Management, Victor Bunderson. Vic shared my enthusiasm for the importance of the project, and blanched only slightly at the stratospheric level of support that I requested to accomplish it. It must not have been easy for him to dig into his discretionary funds to find so many dollars, but he never let on. The resources were found and we were allowed to proceed in as *laissez-faire* a manner as ever I have seen. I am indeed grateful for his trust. I am also delighted that the quality of the final product warrants it.

Next, I would like to thank my colleagues at ETS who provided help and feedback on various aspects of both the NPRDC project and on the writing of this *Primer.* This would be a very long list if I didn't edit it a bit, and so I must mention only some of those who stand out starkly in my memory as especially key. They are:

Eric Perkins and Michael Zieky—who taught me about sensitivity review.

Mari Pearlman and Barbara Foltin—who taught me about computerized test construction.

Carolyn Massad and many members of the ETS Test Development Staff—who taught me what's not in the books about writing items and building tests.

Paul Holland and Charlie Lewis—who continue to teach me about modelling test responses, both with IRT and without it.

Last, my gratitude to my coauthors of this volume. Each of them shouldered the responsibility of writing a section of a book that would integrate with other sections without knowing what those other sections would be like. All of them made major changes in the material that had previously been written for the *CAT-ASVAB Technical Manual,* despite my overly stringent allocation of resources. The final integration of the book is due to cooperation and rewriting that was

truly above and beyond the call of duty. In addition, everyone read and commented on everyone else's work. Such careful editing has resulted in a synergistic effort on our work that has meant producing something better than merely the sum of all of our parts.

Howard Wainer
Princeton, NJ

Preface to the Second Edition

In the decade since we first wrote this *CAT Primer*, we have learned a great deal about CAT. And although there have been important advances in test theory that promise to have an impact on CATs, these are dwarfed by the changes in the delivery system: the world of computing. Computers not only have much more speed and storage than ever before, but these improvements are wrought at considerably lower prices.

The aspect of computing that promises to dominate the future is interconnectivity; computers, and the people who use them, can talk to each other more easily than shouting to the next office. Evidence for this is in the number of us that find that their colleagues send an e-mail rather than walk across the hall. Streaming video, real-time digital audio, and dozens of other miracles are now commonplace. The questions we now must address deal less with "how to use it?" but more often "under what circumstances and for what purposes should we use it?" The future surely holds a promise for the possibilities of testing that are hard to foresee, but tests will still need to fulfill the age-old canons of validity that characterize good practice. Test security remains an essential element for the validity of most tests, and how to maintain security at-a-distance remains an unsolved problem.

A shift in emphasis of the questions asked about CAT has occurred over the past decade, as attempts to make CAT operational have provided data and experience. The importance of the enterprise also has had the effect of increasing the closeness with which those data were scrutinized. This examination revealed practical limitations to the technology that were not apparent earlier. As the glow of initial enthusiasm faded and as our eyes became accustomed to the darker re-

ality, previously unsuspected problems emerged. With our increasing awareness of practical limitations has come the requirement that we reevaluate old assumptions and an accompanying need for a methodology for this evaluation.

The first edition of this book focused on "how to do it." It included some technical advice that subsequent investigations have allowed us to improve upon. This edition has the same focus as the first, but also includes an important caveat in chapter 10. It is important that anyone contemplating the development of a new computer-based test, or the transition of an old one, ask first "Why should I administer this test by computer?" Even with the monumental shrinkage in the costs of computing it remains true that computerized testing is much more expensive than traditional paper-and-pencil tests (see Figure 10.8). If there is no compelling need for what the computer administration offers, it remains sensible to hesitate. Mae West's advice that, "Anything worth doing is worth doing slowly" was wise indeed.

Tests should be computerized if the constructs they are trying to measure cannot be assessed easily without the computer; one example might be tests of architectural design that requires a simulation task embedded within a CAD-CAM environment.

Tests can be computerized if it is important to offer the test continuously in time; examples are licensing tests, where a delay means a loss of income for the successful candidate, and the ASVAB, which historically has been offered continuously.

It is impractical to offer a computerized test in a mass administration a few times a year. Current economic constraints mean that a computerized test must be offered continuously. Continuous testing offers an enormous security challenge when the tests have high stakes for the examinee. This challenge is difficult to meet even with all of the power and flexibility of CAT; it is nigh onto impossible in paper-and-pencil format. We must be sure that we need continuous testing before venturing onto this particular minefield. But if we decide that continuous testing is an important feature (and not an annoying consequence) CAT emerges as a sensible option.

Tests can be computerized if it is important for everyone involved to get the right answer; no sane person would cheat on an eye test. Into this third category falls both diagnostic and placement tests. Moreover, the flexibility of CAT fits very well with the aims of both of these kinds of tests. In diagnostic testing, a CAT can efficiently zero in on exactly what areas are weak. This diagnosis can help guide instruction; when combined with a matched program of instruction it is called a placement test.

High stakes tests whose results are only required once or twice a year are poor candidates for computerized testing; final exams, advanced placement exams, entrance exams all fall into this category.

Chapter 2, which describes the system considerations that are necessary for a CAT is completely new. The computer developments of the past decade made the

earlier version sufficiently archaic that nothing less than a complete rewrite would suffice. I am delighted that Bert Green agreed to do it.

In chapter 4 we have updated the methodology surrounding online calibration to be consonant with current knowledge. We have also included an introduction to the most recent development in the modeling of testlets. This work allows CATs to contain testlets that are constructed on the fly by the item selection algorithm and still estimate the parameters accurately. It does this by modeling the excess local dependence that always seems to manifest itself. I am grateful that Bob Mislevy is enough of a perfectionist to want to take the opportunity to improve our earlier chapter and to urge me to include the hot-off-the-press results that Eric Bradlow, Xiaohui Wang, Zuru Du, and I have recently produced on testlet response theory.

Chapter 10 is new. It contains facts about how items are really selected, what usage patterns emerge, how these patterns influence how many new items are required, and provides some tools for managing item pools. Were it not for the facts and concepts that are required to understand the topic I would have made it the first chapter and not the last. But I urge the reader to stay the course and postpone the decision to build a CAT until after finishing this chapter. I once joked that the information and advice contained in chapter 10 could, if it is followed, keep unwary testing companies out of Chapter 11. I believed it then and I believe it now.

Howard Wainer
Princeton, NJ

This page intentionally left blank

1 Introduction and History

Howard Wainer

PROLOGUE

As we approach the end of the twentieth century we see the influence of computers all around us. In the 1970s computers worked behind the scenes to balance books, write paychecks, prepare weather reports, and do any number of tasks whose characteristics usually included odious repetitive operations. In the 1980s there was a change. Computers came out of the basement. The bank's computer began to deal with the customer first hand, without the human intervention of bank employees. On most desks was a personal computer that processed both words and data, and could be connected with others through telephone networks, which themselves were run by computers. Tasks that computers now do are starting to get more complex. *Machine intelligence, Inference engines,* and *Expert Systems* are terms that are increasingly in vogue.

The use of computers within the context of mental testing has paralleled this development. In the 1970s large testing programs used computers to score tests and process score reports. In the 1980s we have begun to see computers administer exams. The increasingly broad availability of high-powered computing has made possible the administration of types of exam questions that were previously impractical. Moreover, exams could be individualized to suit the person taking them. Of course the development of procedures that adapt to the proficiency of the examinee required the solution of many difficult statistical and psychometric problems. These problems have presented challenges that have only now been solved sufficiently well for practical large-scale application. This volume is a description of how to build, maintain, and use a computerized adaptive testing system (a CAT).

Aristotle, in his *Metaphysics,* pointed out, "We understand best those things we see grow from their very beginnings." We agree. Thus, our description of what we believe is the future of testing begins with a brief glimpse into its past.

THE FIRST FOUR MILLENNIA OF MENTAL TESTING

The use of mental tests appears to be almost as ancient as western civilization. The Bible (Judges 12:4–6) provides an early reference in western culture. It describes a short verbal test that the Gileadites used to uncover the fleeing Ephraimites hiding in their midst. The test was one item long. Candidates had to pronounce the word *shibboleth;* Ephraimites apparently pronounced the initial *sh* as *s*. Although the consequences of this test were quite severe (the banks of the Jordan were strewn with the bodies of the 42,000 who failed), there is no record of any validity study.

Some rudimentary proficiency testing that took place in China around 2200 B.C. predated the biblical program by almost a thousand years. The emperor of China is said to have examined his officials every third year. This set a precedent for periodic exams in China that was to persist for a very long time. In 1115 B.C., at the beginning of the Chan dynasty, formal testing procedures were instituted for candidates for office. Job sample tests were used, with proficiency required in archery, arithmetic, horsemanship, music, writing, and skill in the rites and ceremonies of public and social life.

The Chinese discovered that a relatively small sample of an individual's performance, measured under carefully controlled conditions, could yield an accurate picture of that individual's ability to perform under much broader conditions for a longer period of time. The procedures developed by the Chinese (Têng, 1943) are quite similar to the canons of good testing practice used today. For example, they required objectivity—candidates' names were concealed to insure anonymity; they sometimes went so far as to have the answers redrafted by another individual to hide the handwriting. Tests were often read by two independent examiners, with a third brought in to adjudicate differences. Test conditions were as uniform as could be managed—proctors watched over the exams given in special examination halls that were large permanent structures consisting of hundreds of small cells. Sometimes candidates died during the course of the exams.

This testing program was augmented and modified through the years and has been praised by many western scholars. Voltaire and Quesnay advocated its use in France, where it was adopted in 1791 only to be (temporarily) abolished by Napoleon. It was cited by British reformers as their model for the system set up in 1833 to select trainees for the Indian civil service—the precursor to the British civil service. The success of the British system influenced Senator Charles Sumner and Representative Thomas Jenckes in developing the examination sys-

tem they introduced into Congress in 1868. There was a careful description of the British and Chinese system in Jenckes' report "Civil Service in the United States," which laid the foundation for the establishment of the Civil Service Act passed in January 1883.

Universities lagged far behind in their efforts to install examination systems. The first appears to be the formal exams begun at the University of Bologna in 1219. This was exclusively an oral exam. This structure was also described by Robert de Sorbon, the chaplain of Louis IX, as being used in that court. It was adopted for use in 1257 in the community of scholars that evolved into the Sorbonne. Written tests within universities seem to have their genesis much later with the sixteenth century Jesuits. The first pioneering effort at the development of formal test standards came from this order. In 1599, after several preliminary drafts, eleven rules for the conduct of exams were published. These rules (see McGucken, 1932) are almost indistinguishable from those used today.

The tradition of oral exams spread quickly and by mid-seventeenth century were a standard part of an Oxford education. Written exams were also used and by the middle of the nineteenth century were widely applied in the United States and Western Europe. By the beginning of the twentieth century, serious research efforts had begun on the use and usefulness of various testing procedures. These were done in the United States by Cattell, Farrand (later president of Cornell), Jastrow, Thorndike, Wissler, and Witmer (who founded the first psychological clinic) and in Europe, where Kraepelin (one of Wundt's first students) and Ebbinghaus did important work that eventually led to Binet's intelligence test and Terman's use of it to study "Genius and Stupidity" in his dissertation.

The flurry of activity in testing at the beginning of the twentieth Century spanned a broader range of disciplines than just psychology. One of the most crucial contributions was from statistics, when Spearman provided the rudiments of psychometrics. He invented reliability coefficients and much of the ancillary statistical machinery that allowed their estimation and interpretation.

Tests of all descriptions began to appear to measure performance on such diverse tasks as verbal analogies (devised by Burt, 1911), shoving various shapes through holes (Woodworth, 1910), solving mazes (Porteus, 1915), and drawing a man (Goodenough, 1926). A major change in test administration was occurring at this same time, when there was a shift in practice from individualized to mass administration. This had positive and negative aspects. It allowed much more efficient testing and provided the possibility of a homogeneous testing environment. But it also increased the possibility of examinees not following the directions properly or for some other reason not performing up to their ability.

As the group administered test was evolving, the multiple choice format became increasingly widespread. E. L. Thorndike, at Columbia, and L. L. Thurstone, at Chicago, arranged test material so that items could be scored with a key. Otis, working with Terman at Stanford, was the first to develop an intelligence test that could be scored completely objectively. Prior to the formal

publication of Otis' test, the United States entered World War I; nevertheless Otis' test became the prototype of the *Army Alpha*—the instrument that inaugurated large-scale mental testing.

THE ORIGINS OF MENTAL TESTING IN THE U.S. MILITARY

Robert M. Yerkes, president of the American Psychological Association, took the lead in involving psychologists in the war effort. One major contribution was the implementation of a program for the psychological examination of recruits. Yerkes formed a committee for this purpose which met in May of 1917 at the Vineland Training School. His committee included: W. V. Bingham, H. H. Goddard, T. H. Haines, L. M. Terman, F. L. Wells, and G. M. Whipple. This group debated the relative merits of very brief individual tests versus longer group tests. For reasons of objectivity, uniformity and reliability, they decided to develop a group test of intelligence.

The criteria they adopted (from DuBois, 1970, p. 62) for the development of the new group test were:

1. *Adaptability for group use.*
2. *Correlation with measures of intelligence known to be valid.*
3. *Measurement of a wide range of ability.*
4. *Objectivity of scoring, preferably by stencils.*
5. *Rapidity of scoring.*
6. *Possibility of many alternate forms so as to discourage coaching.*
7. *Unfavorableness of malingering.*
8. *Unfavorableness to cheating.*
9. *Independence of school training.*
10. *Minimum of writing in making responses.*
11. *Material intrinsically interesting.*
12. *Economy of time.*

In just 7 working days they constructed ten subtests with enough items for ten different forms. They then prepared one form for printing and experimental administration. The pilot testing was done with fewer than 500 subjects. These subjects were broadly sampled, coming from such diverse sources as a school for the retarded, a psychopathic hospital, a reformatory, some aviation recruits, some men in an officers' training camp, 60 high school students and 114 Marines at a Navy yard. They also administered either the Stanford-Binet intelligence test

or an abbreviated form of it. The researchers found that their test correlated .9 with the Stanford-Binet and .8 with the abbreviated Binet.

The items and instructions were then edited, time limits revised, and scoring formulas developed to maximize the correlation of the total score with the Binet. Items within each subtest were ordered by difficulty and four alternate forms were prepared for mass administration.

By August, statistical workers under Thorndike's direction had analyzed the results of the revised test after it had been administered to 3,129 soldiers and 372 inmates of institutions for mental defectives. The results prompted Thorndike to call this the "best group test ever devised." It yielded good distributions of scores, correlated about .7 with schooling and .5 with ratings by superior officers. This test was dubbed *Examination a.*

In December of the same year, *Examination a* was revised once again. It became the famous *Army Alpha.* This version had only eight subtests; two of the original ten were dropped because of low correlation with other measures and because they were of inappropriate difficulty. The resulting test (whose components are shown below) bears a remarkable similarity to the cognitive parts of the modern Armed Services Vocational Aptitude Battery (ASVAB), the test currently used by the U.S. armed services.

Test	Number of Items
1. Oral Direction	12
2. Arithmetical Reasoning	20
3. Practical Judgement	16
4. Synonym-Antonym	40
5. Disarranged Sentences	24
6. Number Series Completion	20
7. Analogies	40
8. Information	40

This testing program, which remained under Yerkes' supervision, tested almost 2 million men. Two-third of these received the *Army Alpha,* the remainder were tested with an alternative form, *Army Beta,* a nonverbal form devised for illiterate and non-English-speaking recruits. Together they represented the first large scale use of intelligence testing.

The success of the *Army Alpha* led to the development of a variety of special tests. Link (1919) discovered that a card-sorting test aided in the successful selection of shell inspectors and that a tapping test was valid for gaugers. He pointed out that a job analysis coupled with an experimental administration of

tests thought to require the same abilities as the job and a validity study that correlated test performance with later job success, yielded instruments that could distinguish between job applicants who were good risks and those who were not. Thurstone developed a "rhythm test" that accurately predicted future telegraphers' speed.

Testing programs within the military became much more extensive during World War II. In 1939, a Personnel Testing Service was established in the Office of the Adjutant General of the Army. This gave rise to the *Army General Classification Test* (AGCT) which was an updated version of the *Army Alpha.* The chairman of the committee that oversaw the development of the AGCT was Walter V. Bingham, who served on the 1917 committee that developed *Alpha.* This test eventually developed into a four part exam consisting of tests of (a) reading and vocabulary, (b) arithmetic computation, (c) arithmetic reasoning, and (d) spatial relations. Supplemental tests for mechanical and clerical aptitude, code learning ability, and oral trade were also developed. By the end of the war more than 9 million people had taken the AGCT in one form or another. The Navy and the Army Air Forces participated in the same program, but with some different tests than they required for their own special purposes.

In 1950, the *Armed Forces Classification Test* was instituted to be used as a screening instrument for all services. It was designed to insure appropriate allocation of talent to all branches. This was the precursor of the *Armed Forces Qualification Test* (AFQT) which led in turn to the modern *Armed Services Vocational Aptitude Battery* (the ASVAB).

THE ORIGINS OF ADMISSIONS TESTING FOR AMERICAN UNIVERSITIES

The development of admissions testing at American universities parallel the development of military testing. It was begun in earnest at the beginning of the twentieth century with the founding of the College Board. The first exams were held in June of 1901, at which time 973 candidates wrote essays in one or more of nine subjects: English, French, German, Greek, Latin, history, mathematics, chemistry, and physics. This was hardly a broad sample of examinees because 758 of the 973 were seeking admission to either Columbia or Barnard. But it was a beginning (for a more detailed description of this development the interested reader is referred to Angoff & Dyer, 1971).

By 1925, the success of the Army's testing program had influenced the College Board. An advisory committee was formed whose membership overlapped with Yerkes' 1917 Vineland Committee. This committee was chaired by Carl C. Brigham (who had joined Yerkes' group in October of 1917) and included Yerkes and Henry T. Moore. They recommended the development of a "Scholastic Aptitude Test" to explicitly distinguish it from the achievement

tests then in use. The first SAT was given in June of 1926 to over 8,000 candidates. It was composed of nine subtests and bore a more than passing resemblance to the *Army Alpha*.

Analogies	*Definitions*	*Arithmetical Problems*
Antonyms	*Classification*	*Number Series*
Paragraph Reading	*Artificial Language*	*Logical Inference*

These nine were reduced to seven in 1928 and six in 1929. At about this time Brigham divided the test into two major subsections (one measuring verbal aptitude and the other mathematical) to better suit the different goals of its users.

Until 1937, the SAT was given once a year, in June. But in April of 1937 this changed, when an additional SAT administration was given—principally for scholarship applicants. This Spring administration gained in prominence and it was felt that it would be well if comparisons could be made between the two administrations. This led to the further development and utilization of equating methods and to the definition of a standardization group. All scores were then referred to a group tested in April of 1941 whose mean score was scaled to a mean of 500 and a standard deviation of 100. Subsequent administrations have been equated and scaled to this normative standard.

The use of the test increased in fits and starts, but by the late 1940s it was firmly established and was used to aid in admissions decisions and scholarship competition. The exam reached its current, principally multiple-choice, composition quite early on, for exactly the same reasons that drove the developers of *Alpha*. The costs in time and money for administering any other kind of item were too large for most practical applications.

As the technology for creating valid tests matured, their use broadened to include industrial placement and advancement. Licensing of prospective members of various professions and trades, from actuaries to zoologists, included a *pro forma* standardized test. Increased use marched apace with increased theoretical and technical development.

In 1934, Professor Benjamin Wood at Columbia University joined his staff with engineers from IBM in a collaborative effort to develop a mechanical test-scoring machine. Interestingly, the first workable model was developed by Reynold B. Johnson, a high school science teacher (see Downey, 1965, for a full account of the invention of the first test-scoring machine). His machine used the notion that the number of electrically conductive graphite pencil marks in predetermined positions on a sheet of paper could be reliably read from an ammeter. The invention of this machine had three immediate consequences:

1. It lessened costs by reducing the labor required to grade exams, and by utilizing a separate answer sheet it allowed test booklets to be reused.

2. It stimulated the use of large scale testing programs because mass scoring was now feasible.
3. It increased the reliance on the multiple-choice format for test items.

In 1947, Jane Loevinger stated the concept of *test homogeneity* which would have a profound effect on the future of testing. Loevinger felt that a test should be thought of as a collection of items that were all measuring the same general trait, ability, or function. This idea led to a variety of methods to select items that all measured the same thing. It was also to become the fundamental tenet of item response theory. In a sense, her proposal of homogeneity was a reaction to the epistemological difficulties raised by the findings of factor analysts who uncovered the multiplicity of underlying skills needed to correctly answer many of the existing tests. These factor analyses gave rise to Thurstone's well-known *Primary Mental Abilities,* as well as Guilford's much more molecular mental factors.

The first major compendium of formal psychometric methods specifically designed to construct, score, and interpret ability/proficiency tests was written by Harold Gulliksen of Princeton University and the Educational Testing Service, and appeared in 1950. A year later, John Flanagan (1951) proposed a formalization of existing procedures for test construction. He suggested the use of *item rationales* to construct new tests. This involves first listing the behaviors that are to be tested. Specifications are then prepared for the items whose purpose is to measure each of these behaviors. This systematic approach replaced the more informal procedures that had been in general use previously.

A capstone was placed on traditional test theory in 1968 with the publication of Lord and Novick's *Statistical Theories of Mental Test Scores.* It simultaneously accomplished three things:

1. It summarized all of the important work in test theory up until that time in a cohesive way.
2. It provided a formal mathematical structure to support the various aspects of traditional test theory (*true score theory*). In so doing, the assumptions and axioms of true score theory were made explicit. This clearly showed the strengths and weaknesses of existing theory while providing the statistical machinery to best exploit the former and to remedy the latter.
3. It introduced the work of the statistician Allan Birnbaum to the psychometric literature. Birnbaum's five chapters in Lord and Novick provide the basis of modern item response theory (IRT). In it he leans on insights (like Loevinger's homogeneity idea) that underlay traditional true score theory, as well as earlier work on latent trait models (e.g. Rasch, 1960). But he went much further, providing the statistical foundations of a test theory that considers the item, rather than the entire test, as its fundamental unit.

> This formal theory clarified many issues and allowed the graceful solution to many problems that previously were dealt with in a much clumsier way.

Although IRT had many obvious advantages, its real strength was that it could deal with items one-at-a-time. It posited an underlying, unobserved trait, on which the items were linearly arrayed from the easiest to the hardest. The goal of testing was to be able to array the examinees on the same continuum as the items, from novice to expert. This goal meant that one did not have to present all items to all individuals, only enough items to allow us to accurately situate an examinee on the latent continuum. The power to do this did not exist comfortably within the confines of traditional true score theory and yet was a natural outgrowth of IRT. In fact, the capacity to rank all examinees on the same continuum, ***even if they had not been presented any items in common,*** gave rise to the possibility of a test that was individually tailored to each examinee. Such a test is called *Adaptive,* and many believe that adaptive testing is the *raison d'etre* of IRT.

COMPUTERIZED ADAPTIVE TESTING

Throughout its entire history there has always been the tradeoff between individual testing and group testing. An individually administered test does not contain too many inappropriately chosen items and, furthermore, we are assured that the examinee understands the task. A group-administered test has the advantage of uniformity of situation for all examinees, as well as a vastly reduced cost of testing. Throughout this century, the choice has almost always been in favor of the mass-administered test.

A critical problem facing a mass-administered test is that it must be assumed that there is a relatively broad range of ability to be tested. To effectively measure everyone, the test must contain items whose difficulties match this range (i.e., some easy items for the less proficient, some difficult ones for the more proficient). If the test did not have difficult items, we might not, for example, be able to distinguish among the proficient examinees who got all the easy items correct. Similarly, if there were no very easy items on the test, we might not be able to distinguish among the less proficient examinees who got the more moderate items all wrong. If making these kinds of discriminations is important, the test must contain as broad a range of item difficulties as the proficiency range of the population to be tested. The accuracy with which a test measures at any particular proficiency level is (roughly) proportional to the number of items whose difficulties match that level.

Fortunately for mass-administered testing, Lincoln's observation that "the good Lord must have loved the common man because he made so many of

them'' remains valid. Most examinees' abilities seem to lie in the middle of the continuum. Thus, mass tests match this by having most of their items of moderate difficulty with fewer items at the extremes.

The consequence of this test structure has historically been that the most proficient examinees have had to wade through substantial numbers of too easy items before reaching any that provided substantial amounts of information about their ability. This was wasteful of time and effort as well as introducing possibly extraneous variables into the measurement process, for instance, the chance of careless errors induced by boredom. Less proficient examinees face a different problem. For them, the easy items provide a reasonable test of ability, whereas the difficult ones yield little information to the examiner. They can, however, cause confusion, bewilderment, and frustration to the examinee. They also add the possibility of guessing, which injects extraneous noise into the measurement process.

In the early 1970s, the possibility of a flexible mass-administered test that would alleviate these problems began to suggest itself. The pioneering work of Frederic Lord (1970, 1971a,b,c,d) is of particular importance. He worked out both the theoretical structure of a mass-administered, but individually tailored test, as well as many of the practical details.

The basic notion of an adaptive test is to mimic automatically what a wise examiner would do. Specifically, if an examiner asked a question that turned out to be too difficult for the examinee, the next question asked would be considerably easier. This stems from the observation that we learn little about an individual's ability if we persist in asking questions that are far too difficult or far too easy for that individual. We learn the most when we accurately direct our questions at the same level as the examinee's proficiency. An adaptive tests first asks a question in the middle of the prospective ability range. If it is answered correctly, the next question asked is more difficult. If it is incorrectly answered, the next one is easier. This continues until we have established the examinee's proficiency to within some predetermined level of accuracy.

Early attempts to implement adaptive tests were clumsy and/or expensive. The military, through various agents (e.g., Office of Naval Research; Navy Personnel Research and Development Center; Air Force Human Resources Laboratory; Army Research Institute) recognized early on the potential benefits of adaptive testing and supported extensive theoretical research efforts. Through this process much of the psychometric machinery needed for adaptive testing was built. Nevertheless, the first real opportunity to try this out in a serious way awaited the availability of cheap, high-powered computing. The 1980s saw this and the program to develop and implement a computerized adaptive test (CAT) began in earnest.

This work was aimed at improving the entire measurement process. In addition to the increased efficiency of testing the other advantages of a CAT (from Green, 1983) are:

1. Test security is improved, to the extent that a test is safer in a computer than in a desk drawer. Moreover, because what is contained in the computer is the item pool, rather than merely those specific items that will make up the examinee's test, it is more difficult to artificially boost one's score by merely learning a few items. This is analogous to making available a dictionary to a student prior to a spelling test and saying, "All the items of the test are in here." If the student can learn all of the items, the student's score is well earned.
2. Individual's can work at their own pace, and the speed of response can be used as additional information in assessing proficiency. Aside from the practical necessity of having rough limits on the time of testing (even testing centers must close up and clean the floors occasionally), we can allow for a much wider range of response styles than is practical with traditional standardized tests.
3. Each individual stays busy productively—everyone is challenged but not discouraged. Most items are focused at an appropriate range of difficulty for each individual examinee.
4. The physical problems of answer sheets are solved. No longer would a person's score be compromised because the truck carrying the answer sheets overturned in a flash flood—or other such calamity. There is no ambiguity about erasures, no problems with response alternatives being marked unwittingly.
5. The test can be scored immediately, providing immediate feedback for the student. This has profound implications for using tests diagnostically.
6. Pretesting items can be easily accomplished by having the computer slip new items unobtrusively into the sequence. Methods for doing this most effectively are still under development, but see chapter 4 for one method.
7. Faulty items can be immediately expunged, and an allowance for examinee questioning can be made.
8. A greater variety of questions can be included in the test builder's kit. The multiple-choice format need not be adhered to completely—numerical answers to arithmetic problems can just be typed in. Memory can be tested by use of successive frames. With voice synthesizers, we can include a spelling test, as well as aural comprehension of spoken language. Video disks showing situations can replace long-winded explanations on police or firefighter exams.

IMPORTANT ISSUES IN CAT

This area is dealt with in greater detail in subsequent chapters, however we give a flavor of some of them here.

Psychometric Theory

Different examinees taking a CAT, in all likelihood, take different forms of the test. A very proficient examinee might have few (or even no) items in common with someone who was considerably less proficient. This never happened with traditional tests in which everyone had the same items. In a traditional test, a measure like "number correct" worked fine. In a CAT that would not work, because (if the test is working properly) all examinees would get about half of the items presented to them correct. The more proficient examinees would get half of a rather difficult subset correct. The less proficient would get their half out of a much easier subset. The glue that holds all of the different tests together is a particular kind of psychometric theory called Item Response Theory (IRT)—see Wainer, (1983) for a particularly readable account of this complex statistical theory; chapters 3 and 4 contain details and further references.

Briefly, IRT presents a mathematical characterization of what happens when an individual meets an item. Each individual is characterized by a proficiency parameter (usually denoted θ) and each item by a collection of parameters—one of which is the item's difficulty (here denoted b). The IRT model compares the person's proficiency with the item's difficulty and predicts the probability of that person getting that item correct. If the person is much more proficient than the item is difficult, then this probability will be large. If the item is much more difficult than the person is proficient, then this probability will be small. We learn the most when this expected probability is close to one-half ($p = .5$). The item-choice algorithm tries to pick items that yield the greatest amount of information while at the same time satisfying the variety of content specifications that are critical for a good test. An examinee's proficiency is calculated from the difficulty of the items that are presented to him.

System Design and Operations

In a paper-and-pencil (P&P) test administration certain standards must be maintained. Rooms where the tests are administered have to have desks and chairs suitably spaced and configured so that examinees can be fairly measured. Lighting must be sufficient so that test forms can be read easily. Temperature must be controlled so that examines are comfortable. In general, care must be exercised to prevent compromising the validity of the test in all of its aspects. Identical concerns exist within the context of a CAT. But some of these concerns show up in different ways. We must control glare on the screen. We must worry more about system reliability and backup systems (this is analogous to keeping a box of extra #2 pencils on hand for the P&P version, but a good deal more complex). In a CAT, we must be sure that displays have adequate resolution for both graphics and text; that branching processes work properly; that item presentation

and test scoring software works impeccably. Bert Green provides a detailed description of these issues in chapter 2.

Item Pool Development and Testing

The building blocks out of which a test is constructed are its component items. If they are not well constructed no statistical magic nor electronic wizardry will help. Issues of item pool construction are discussed by Ronald Flaugher in chapter 3, along with the methodology of item construction, pretesting, and screening.

The process of item pool development is a long and arduous one. This is seen in stark contrast to the 6 month item development process of the *Army Alpha*. The reasons for this are several.

First, the state-of-the-art of test development has advanced considerably; many issues are now important which were not thought of 70 years ago. For example, careful consideration is given to item content as it bears on the depiction of women and minorities. It is well established that sensitivity to issues of this nature has yielded tests with broader validity than earlier tests.

Secondly, a CAT makes much more stringent demands on its component items than does its paper-and-pencil counterpart. Because the CAT tends to be much shorter (in general a CAT is about half as long as a traditional test yielding about the same accuracy of measurement), each item is more critical. If an item is flawed, its impact on the estimate of the examinee's proficiency is doubled. Additionally, because not everyone gets the same set of items, a flawed item can affect some examinees and not others. Hence, test fairness, in addition to test validity, can be compromised.

Chapters 3, 7, and 8 describe the careful processes of item pool development and checking that are necessary to assure that the items in a CAT are as flawless as can be made. These processes include screening by a panel of subject matter experts, a sensitivity review panel, and a group of test development experts. This is in addition to extensive pretesting that includes validity and reliability studies.

Item Response Theory

Chapter 4 introduces item response theory in some detail. This is the theoretical glue that holds a CAT together. In this chapter Robert Mislevy and I provide both the logic that was the genesis of IRT and the equations that are its manifestation. We also describe the details involved in calibrating an item pool and in scoring a test. This chapter is a bit heavy mathematically, but this cannot be avoided if we are to provide precision in our prescriptions.

Testing Strategies and Choices

The key questions in a CAT are:

1. How do we choose an item to start the test?
2. How do we choose the next item to be administered after we have seen the examinee's response to the current one?
3. How do we know when to stop?

The concept of Test Information provides the start of an answer. Test Information is (roughly) the inverse of the variance of estimation. Thus, if we have large error bounds surrounding an estimate of an examinee's ability, we have little information. One notion of an item choice algorithm is, at every stage of the test, to choose that item that yields the largest marginal gain to the information we have. This can even work for the first item; we must merely assume some distribution of ability for the examinee population and the optimal item pops out. A stopping rule is also suggested; keep testing until the examinee's ability has been measured to a preestablished level of accuracy.

In practice it is not that simple. We cannot always begin the test with the same item, because pretty soon everyone would know the answer to that item. We must introduce some variability.

Choosing the maximally informative item as the next one is also a pretty idea, but it does not work in practice. We must be sure to ask questions that cover the content specifications. For example, in an arithmetic examination we might find that the most informative next item tests multiplication, but we have already given plenty of multiplication items and instead need to cover fractions. Thus, we must ask the most informative fraction item.

Stopping only when we have a sufficiently accurate estimate is also attractive, yet how long can we afford to tie up a machine and an examinee? At one point or another practical concerns may force us to stop the process even if acceptable error bounds have not yet been reached.

David Thissen and Robert Mislevy provide a much more complete and technically accurate explanation of the testing algorithms that may be employed in a CAT in chapter 5.

Test Equating

One of the key elements of a scientific system of measurement is its comparability across different places and different times. How much would a measure of weight be worth if the scale at the doctor's office had no relation to the scale in your bathroom? The two measuring instruments must be either equivalent (the weight on one is the same, within the margin of error, as on the other)

or equatable (the weight on one, perhaps measured in kilograms, may be equated to the weight on the other, which perhaps was measured in pounds).

The identical problem surfaces in mental testing. A score that establishes proficiency in a particular skill this year needs to hold up next year as well. A test given to one person should be comparable to another given to someone else. We must be able to equate the many different forms of the test to one another.

In addition to these kinds of problems, large testing programs contemplating moving to a CAT format encounter transition problems. During the time of switchover from a P&P to a CAT, no examinee should be at a disadvantage simply because of the test format assigned to him. Thus, the forms must be carefully equated so that an examinee (who knew all of the details and ramifications) would be indifferent as to which form she was assigned.

The procedures involved in both equating tasks—equating P&P to CAT and equating one CAT form to another—are complex. They are described by Neil Dorans in chapter 6.

Reliability and Precision

Reliability refers to the degree to which a test is free from error. Reliability in tests is as important as reliability in machinery. The formal concept of reliability within mental tests is commonly credited to Spearman, yet Edgeworth made important contributions to the theory of reliability in the scoring of essays more than two decades earlier (1888, 1892). This concept, whoever invented it, basically involves the notion of ranking a bunch of people on their performance on a test and then reranking them based on another form of the same test. The extent to which the examinees maintain the same order in both rankings reflects the reliability of the test.

The ranking concept is useful, but has several shortcomings. For example, it is highly dependent upon the distribution of ability of the examinees sampled. If they are very homogeneous (all the same) any test will look unreliable. If they are very diverse a test may look wonderful. It was surely no accident that the developers of *Alpha* always included a school for the feebleminded in their norming sample. Doing this widened the ability distribution and hence showed the performance of the new test in a very favorable light.

But the should the precision of a measuring instrument depend on who is being measured at the same time? Should your weight depend on who used the scale before you? Of course not. Modern IRT provides an alternative to traditional reliability—the standard error of the ability estimate. Rather than saying that the test's reliability is .86 or something, we instead now state that someone's proficiency θ is equal to $\hat{\theta} \pm \varepsilon$. When ε is sufficiently small, we are satisfied. This idea of standard error is key to the notion of measurement precision and is one of the major advances that IRT has allowed.

David Thissen provides the details behind these concepts in chapter 7.

Validity

"Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness and usefulness of *the specific inferences made from test scores.*" These strong words, stated in the *Joint Technical Standards for Educational and Psychological Testing,* underscore the importance of the concept of validity as the touchstone of testing. Note that it is not the test that has validity, but rather *the inferences* made from the tests scores. Thus, before we can assess a test's validity, we must know the purposes to which it is to be put.

The validity of a CAT can be compromised in one or more of its three component parts. These areas are:

1. The validity of the items in predicting performance. If too many of the items are flawed, the resulting scores may be meaningless.
2. The validity of the methodology of computer presentation. If the fact that items are presented on a computer screen, rather than on a piece of paper, changes the mental processes requires to respond correctly to the item, the validity of inferences based on these scores may be changed.
3. The validity of the item selection algorithm. If the item selection algorithm does not make up tests wisely (i.e., not spanning the content specifications evenly) we might find that the validity of inferences based on these scores is threatened.

The validity of a new CAT may be tested by comparing it to the P&P test that preceded it and whose validity has been previously studied. By showing that the new CAT can be successfully equated to the P&P version, we provide prima facie evidence for its validity. Other studies are still required, for example, studies that compare the validity across a variety of demographic subgroups. Most important are studies that look at the three areas that set a CAT apart from previous tests. These studies are described in greater detail by Lynne Steinberg, Thissen, and me in chapter 8.

CHALLENGES FOR THE FUTURE

Adaptive tests are only the beginning. As we outlined earlier, the crucial letter in CAT is *C*. Computerizing test administration opens the door for many new kinds of tests. We need not concern ourselves with speeded tests; instead, we can measure how fast an examinee answers a question. Indeed, we can compare speed at answering different parts of questions. We can, at last, abandon the multiple-choice format for many kinds of questions. Instead, we can require the examinee to answer the question directly. With voice synthesizers, we can have a

spelling test. Other complex kinds of tests can be developed. Driving simulators may serve as better predictors of success on the road than paper-and-pencil tests.

All of this is in addition to the possibility of increasing the amount of testing in an optimal way, without wasting time asking too hard or otherwise inappropriate questions.

It is a new world, and with careful scholarship and creative thoughts, the resulting testing program should allow greater utilization of human talent than ever before. And at a cost that would allow its use on an ever-broadening scale. Yet, all the problems are far from being solved. There are many challenges to be faced. Can we count on the calibration of items to remain stable over time? Over changes in examinee population? Over differences in context? How can we assure ourselves that Flanagan's concerns about content balance are satisfied? Do high-scoring examinees get a test on the same subject as low-scoring examinees? On a broad scale mathematics test, can we write easy calculus items and hard arithmetic items? Need we worry about such things? What happens when the IRT model we normally use is inappropriate? In chapter 9, Neil Dorans, Bert Green, Robert Mislevy, Lynne Steinberg, David Thissen and I discuss the exciting challenges that still lie before us.

THE STRUCTURE AND USE OF A GEDANKEN COMPUTERIZED ADAPTIVE TEST (THE GCAT)

From the beginning of our efforts in the writing of this *Primer*, we felt that including a single unifying example throughout the *Primer* would be advantageous for two reasons. It would allow us to concretely illustrate many of the issues that were important for us to discuss and by making those issues concrete, ease the comprehension problems of the prospective reader. But what was the right example? We needed one that was clear and simple while simultaneously being rich enough and deep enough to contain all of the areas of discussion. The best solution to this was to makeup a hypothetical exam that would serve our purposes precisely. Since this *CAT is wholly hypothetical* we have dubbed it the

Gedanken Computerized Adaptive Test,

or ***GCAT*** for short. Although it is not a real operational test, it shares its characteristics with many real applications. Among the currently operational precursors to the GCAT are the Army's CAST system, the College Board's CAT placement tests, Lord's wide-range vocabulary test, the Psychological Corporation's CAT version of their Differential Aptitude Battery, and Assessment Systems' package of CAT software.

Additionally, anyone familiar with CAT-ASVAB (and its accelerated version, ACAP) will see many similarities. These are not accidental. Even though

neither of these programs are yet fully operational, their development has incorporated much of the best judgment of the foremost workers in the field. Consequently, we have borrowed from them.

There are many similarities among these various CAT applications which are shared by GCAT. Nevertheless, we hold none of our predecessors responsible for any errors in our formulation of the GCAT, although they certainly should share responsibility for whatever merits it may contain.

We often use the GCAT to illustrate the concept of interest in the specific; we then go on to describe in general the various alternatives and variations on that theme. Whenever we discuss the hypothetical GCAT we typographically set it off in shaded boxes. This is so that no one confuses this exam with any real operational exam.

Background

A large Midwestern state gives a general battery of tests to aid them in selection and placement decisions. Annually, they have more than 65,000 job openings of a wide variety of sorts at many levels. Among these are: entry level clerical positions, management trainee positions, state police, forest rangers, actuaries, and so on. Typically, there are three to four times as many applicants as there are available positions, although some openings are more popular than others. All applicants take one form or another of the test battery, although not all tests are included in all forms. After taking this battery, successful candidates for some positions are offered a job, for other positions they are offered entry to a training program, in still others they are channeled to other placement tests.

Earlier incarnations of this same battery have been in use, in paper-and-pencil format, for more than 30 years. For reasons of efficiency, test security, and for the ability to utilize new tests that cannot be easily administered without a computer, they are in the process of being switched to a CAT format.

The GCAT Battery

The GCAT Battery is made up of six tests. Its makeup is reminiscent of the Army Alpha, the ASVAB, the early SAT, and many other tests that are either in use or have been used. Its six tests are: Vocabulary, Quantitative Reasoning, Science Knowledge, Paragraph Comprehension, Spatial Memory, and Clerical Speed.

The authors of the rest of this *Primer* use and amplify this example to illustrate their explanations of the operation of a CAT. It must be remembered that this is but one example, and that the technology of computerized adaptive testing is much broader than can be encompassed in a single example. Thus we will also

endeavor to sketch some of the directions for such expansion without fully illustrating them. the prospective impact of computerized test administration for increasing the validity and hence usefulness of tests may be as great as was the use of standardized objectively scored tests. But before this is known, there is much work to be done. We need to meld the imagination of test developers to the technology of modern computing. This *Primer* is an attempt to chronicle the initial efforts of that joining.

ACKNOWLEDGMENTS

I am grateful for helpful comments on an earlier draft by Bert F. Green, Jr., Paul W. Holland, Charles Lewis, Martha Farnsworth Riche, and David Thissen. All flaws or errors that remain are my own.

ANNOTATED REFERENCES

Angoff, W. H., & Dyer, H. S. (1971). The admissions testing program. In W. H. Angoff (Ed.), *The College Board Admissions Testing Program* (pp. 1–13). New York: College Entrance Examination Board. A technical manual for the College Board's SAT. It contains many details for one of the most widely known exams given in the United States, as well as some fascinating history.

Burt, C. (1911). Experimental tests of higher mental processes and their relation to general intelligence. *Journal of Experimental Pedagogy, 1,* 93–112. An early British psychometrician describing some of his new intelligence tests.

Downey, M. T. (1965). *Ben T. Wood, educational reformer.* Princeton, NJ: Educational Testing Service. Amidst its biographical mission, this contains the surprising details surrounding the development of the first mechanical test scorer.

DuBois, P. H. (1970). *A history of psychological testing.* Boston: Allyn & Bacon. A complete history of testing that is far broader than the brief sketch we present here. Indeed much that is here was taken from this source.

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51,* 599–635. This paper (and the next one on this list) describe the British Statistician Edgeworth's contributions to test theory. Specifically the development of reliability and its surrounding concepts.

Edgeworth, F. Y. (1892). Correlated averages. *Philosophical Magazine,* 5th series, *34,* 190–204.

Flanagan, J. C. (1951). The use of comprehensive rationales in test development. *Educational and Psychological Measurement, 11,* 151–155. A description of a formal way of utilizing test specifications in test construction. This was one forerunner of the formal concept of content validity.

Goodenough, F. L. (1926). Measurement of intelligence by drawings. Yonkers: World Book. One of many attempts to measure intelligence with nonlinguistic tasks; in this case drawing pictures.

Green, B. F., Jr. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of Modern Psychological Measurement.* (pp. 69–80). Hillsdale, NJ: Lawrence Erlbaum Associates. An expository article on the promise of CAT in the book whose title is almost always misspelled.

Gulliksen, H. O. (1950). *A theory of mental tests,* New York: John Wiley & Sons; Reprint 1987,

Hillsdale, NJ: Lawrence Erlbaum Associates. At the time of its publication this was the most comprehensive statement of mental test theory, and it remained that for almost 20 years.

Jenckes, T. A. Civil Service of the United States. Report No. 47, 40th Congress, 2nd Session, May 25, 1868. The report that led to the development of the modern U.S. Civil Service System. It cites, as its model, the Chinese system, pointing toward its three millennia of existence as a measure of its validity and success.

Link, H. C. (1919). *Employment psychology*. New York: Macmillan. A full explanation of a program of research on job skill testing. It includes job analysis, test development and validity studies, the results of the latter were used to modify the test.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability, *Psychological Monographs, 61*, 4. A clear statement of how all items on a test ought to be chosen so that they all measure the same underlying ability or trait. This is the fundamental tenet underlying modern IRT.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance*, pp. 139–183. New York: Harper and Row. The initial statement on adaptive testing from its progenitor. This chapter, combined with his four 1971 papers, forms the theoretical and psychometric basis for adaptive testing.

Lord, F. M. (1971a). The theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement, 31*, 805–813.

Lord, F. M. (1971b). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*, 147–151.

Lord, F. M. (1971c). Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association, 66*, 707–711.

Lord, F. M. (1971d). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement, 31*, 3–31.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley. The bible of modern test theory. It places a capstone on traditional true score theory and provides a thorough introduction to the formal statistical models of modern item response theory.

McGucken, W. J. (1932). *The Jesuits and education*. Milwaukee: Bruce Publishing. Provides a scholarly look at the role the Jesuits played in the development and use of tests; also discusses their influence on the rest of education.

Porteus, S. D. (1915). Mental tests for the feebleminded: A new series, *Journal of Psycho-Asthenics, 19*, 200–213. Early development of a nonverbal test. It involved the examinee working through a printed maze with a pencil. It was viewed as a supplement to the Binet in the determination of mental retardation.

Rasch, G. (1960). *Probability models for some intelligence and attainment tests*. Copenhagen: Nielsen and Lydiche. A full and lucid account of the simplest of the item response models by its inventor.

Têng, Ssu-yü. (1943). Chinese influence on the western examination system. *Harvard Journal of Asiatic Studies, 7*, 267–312. The source of much of what English readers know about the tradition of exams in China.

Wainer, H. (1983). On item response theory and Computerized Adaptive Tests: The coming technological revolution in testing. *The Journal of College Admissions, 28*, 9–16. A nontechnical description of item response theory and its role in adaptive testing.

Woodworth, R. S. (1910). Race differences in mental traits, *Science, 31*, 171–186. A description of how the author used the Seguin Form Board (a test that requires the examinee to put star-shaped blocks in star-shaped holes, round blocks in round holes, square blocks in square holes, etc.) to test immigrants at Ellis Island for mental defects.

EXERCISES/STUDY QUESTIONS

1. How long has formalized testing been taking place?
2. Why was the tradition of testing in Western Europe largely oral until after the Crusades?
3. What was the historical model on which the American Civil Service System based?
4. When was testing begun in the U.S. military? Why?
5. Why did the initial SAT resemble the test used in the military?
6. What were the key technological events that allowed the development of adaptive testing?
7. What are the principal advantages of CAT?

This page intentionally left blank

References

2 System Design and Operation

u.s. Department of Defense. (1986). A review of the development and implementation of the ASVAB

Forms II. 12. & 13. Washington, DC: Author.

Kreitzberg, C. B., & Jones D. H. (1980). An empirical study of the broad-range tailored test of verbal

ability (RR-80-5). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1977). A broad-range test of verbal ability. Applied Psychological Measurement. 1,

95-100.

McBride, 1. R. (1988, August). A computerized adaptive version of the Psychological Corporation's

Differential Aptitude Battery. Paper presented at the annual meeting of the American Psychologi

cal Association, Atlanta, GA.

McBride, 1. R. & Sympson, J. B. (1985). The computerized adaptive testing system development proj

ect. (pp. 342-349). In D. 1. Weiss, (Ed.), Proceedings of the 1982 Item response theory and Com

34 GREEN puterized Adaptive Testing Conference. Minneapolis: Department of Psychology, University of Minnesota.

Moreno, K. E, Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery and computerized adaptive testing subtests. Applied Psychological Measurement. 8,155-163.

Rafacz, B., & Hetter, R. D. (1997). ACAP hardware selection, software development and acceptance testing. In W. A. Sands, B. Waters, & 1. R. McBride (Eds.), Computerized adaptive testing: From inquiry to operation (pp. 145-156). Washington, DC: American Psychological Association.

Rafacz, B. A., & Tiggle, R. B (1985). Interactive screen dialogues Jor the examinee testing (ET) station developed in support oJthe accelerated CAT-ASVAB project (ACAP). San Diego, CA: Navy Personnel Research and Development Center.

Reed, A. V. (1979). Microcomputer display timing: Problems and solutions. Behavior Research Methods and Instrumentation. 11,572-576.

Sands, W. A., & Gade, P. A. (1983). An application of computerized adaptive testing in Army recruiting. Journal oJComputer-Based Instruction. 10. 37-89.

Sands, W. A., Waters, B, & McBride, 1. R. (Eds.). (1997). Computerized adaptive testing: From inquiry to operation. Washington, DC: American Psychological Association.

Vale, C. D. (1981). Implementing the computerized adaptive test: What the computer can do for you. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. Machine-Mediated Learning. 2, 217-282.

Weiss, D. J. (1974). Strategies oj adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Psychometric Methods Program.

Weiss, D. 1. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492. 2. SYSTEM DESIGN AND OPERATION 35 EXERCISES/STUDY QUESTIONS

1. A CAT system can be part of a network, or a stand-alone system. What are the advantages and disadvantages of each approach?

2. Explain the different functions performed by the ET and TA stations in a networked system.

3. Elaborate the issues involved in deciding how portable to make the system.

4. Discuss the importance of maintaining equivalence among different kinds of testing stations.

5. What precautions should be taken to insure that the room

in which a CAT is given is adequate to the task?

6. How is test security implemented in a CAT system? This page intentionally left blank

3 Item Pools

American Psychological Association. (1985). Standards for educational and psychological testing.

Washington, DC: Author.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. Applied

Psychological Measurement. 12. 261-280.

College Entrance Examination Board. (1981). An SAT: Test and technical data for the Scholastic

Aptitude Test administered in April 1981. New York.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). Manual for kit of factor-referenced

cognitive tests. Princeton, NJ: Educational Testing Service.

Greaud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of

speed tests. Applied Psychological Measurement. /0. 23-34.

Hardwicke, S. B., & Yoes, M. E. (1984). Attitudes and performance on computerized vs. paper

and-pencil tests. San Diego, CA: Rehab Group.

Hunter, R. V., & Slaughter, C. D. (1980). ETS test sensitivity review process. Princeton, NJ:

Educational Testing Service.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: an exploratory

IRT model fit tool. Applied Psychological Measurement, 9, 281-288.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NI: Lawrence Erlbaum Associates.

Prestwood, I. S., Vale, C. D., Massey, R. H., & Welsh, I. R. (1985). Armed service vocational aptitude battery: Development of an adaptive item pool. (Technical report 85-19). San Antonio, TX: Air Force Human Resources Laboratory.

Segall, D. O. (1987). ACAP item pools: Analysis and recommendations (Draft technical report). San Diego, CA: Navy Personnel Research and Development Center.

Thissen, D. M., & Steinberg, L. (1984). A response model for mUltiple choice items. Psychometrilro. 49. 501-519.

Thissen, D. M., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple choice models: The distractors are also part of the item. Journal of Educational Measurement. 26. 161-176.

Wainer, H. (1989). The future of item analysis. Journal of Educational Measurement. 26. 191-208.

Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. Machine-Mediated Learning. 2. 217-282.

Wesman. A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), Educational Measurement (2nd Edition). American Council on Education: Washington, D.C.

Wilbur, E. R. (1986). Design and development of the ACAP test item data base. Proceedings of the 28th annual conference of the Military Testing Association (pp. 601-605). Mystic, CT: U.S. Coast Guard. 3. ITEM POOLS 59
EXERCISES/STUDY QUESTIONS

I. How does a CAT "item pool" serve the same purpose as several forms of a fixed format test? How is it different?

2. How is the process of constructing a CAT item pool different from that of building a fixed format test?

3. Why is item quality control even more important in a CAT than in a fixed format test?

4. What are some reasons for rejecting a candidate item from the CAT pool?

5. When a CAT is replacing an existing paper-and-pencil test, what are some of the issues that must be addressed?

6. How does sensitivity review affect statistical measures of item quality?

7. What is multidimensionality? How does its existence threaten the validity of a CAT? This page intentionally left blank

4 Item Response Theory, Item Calibration, and Proficiency Estimation

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling.

Journal of Educational Statistics, 17, 251-269.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An

application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full infonnation item factor analysis. Applied Psychological Measurement, 12, 261-280.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. Psychometrika, 64, 153-168.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38.

Divgi, D. R., & Stoloff, P. H. (1986). Effect of the medium of administration on ASv.4B item response curves (CNA 86-24). Alexandria, VA: Center for Naval Analysis.

Gibbons, R. D., Bock, R. D., & Hedeker, D. (1987, June). Approximating multivariate normal orthant probabilities using the Clark algorithm. Paper presented at the annual meeting of the Psychometric Society, Montreal.

Green, B. F., Bock, R. D., Linn, R. L., Lord, F. M., & Reckase, M. D. (1983). A plan for scaling the Computerized Adaptive ASVAB. Baltimore, MD: Department of Psychology, Johns Hopkins University.

Hardwicke, S. B., & White, K. D. (1983). Predictive utility evaluation of computerized adaptive testing: Results of the Navy research. San Diego, CA: Rehab Group.

Levine, M. (1985). The trait in latent trait theory. In D. 1. Weiss (Ed.), Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference (pp. 41-65). Minneapolis, MN: Computerized Adaptive Testing

Laboratory, Department of Psychology, University of Minnesota.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: AddisonWesley.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49,359-381.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. Psychometrika, 51, 177-195.

Mislevy, R. J., & Bock, R. D. (1993). BILOG 3.04: Multiple-group lRT Analysis and Test Maintenance for Binary ltems [computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M.S. (1993). How to equate tests with little or no data. Journal of Educational Measurement, 30, 55-78.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOG 1ST and BILOG. Applied Psychological Measurement, 13, 57-75.

Mislevy, R. J., Wingersky, M. S., & Kingston, M. (1990). Evaluation of a procedure for calibrating "seeded" test items. Final report to Battelle Coumbus Division, Contract No. DAAL03-86-D-000I, Delivery Order 0708, Scientific Services Program. Princeton, NJ: Educational Testing Service.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Denmarks Paedagogiske Institut.

Rosenbaum, P. R. (1988). A note on item bundles. Psychometrika, 53, 349-360.

Samejima, F. (1979). A new family of models for the multiple choice item (Research Report #79-4). Department of Psychology, University of Tennessee.

Samejima, F. (1983). Some methods and approaches of estimating the operating characteristics of discrete item responses. In H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement (pp.159-182). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stocking, M. L., & Lord, F. M. (1983). Developing a common

metric in item response theory. Applied Psychological Measurement, 7,201-210.

Sympson, 1. B. (1983, June). A new IRT model for calibrating multiple choice items. Paper presented at the annual meeting of the Psychometric Society, Los Angeles.

Thissen, D. (1986). Multilog: A user's guide. Mooresville, IN: Scientific Software.

Thissen, D., Steinberg, L. & Mooney, 1. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. Journal of Educational Measurement, 26, 247-260.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), Test Validity (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.
4. ITEM RESPONSE THEORY, CALIBRATION, AND ESTIMATION 99

Wainer, H., Bradlow, E., & Du, Z. (2000). Testlet response Theory: An analog for the 3-PL useful in

adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), Computerized adaptive testing:

Theory and practice. Boston, MA: Kluwer-Nijhoff.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets.

Journal of Educational Measurement. 24, 185-202.

Wainer, H., Morgan, A., & Gustafsson, J-E. (1980). A review of estimation procedures for the Rasch

model with an eye toward longish tests. Journal of Educational Statistics, 5, 35~.

Wolfe, 1. H., Moreno, K. E., & Segall, D. O. (1997). Evaluating the predictive validity of the CAT

ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing:

From inquiry to operation (pp. 175-180). Washington, DC: American Psychological Association.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (1993). BIMAIN 2: Item analysis and test

scoring with binary logistic models [computer program]. Mooresville, IN: Scientific Software.

100 WAINER AND MISLEVY EXERCISES/STUDY QUESTIONS 1. What is Item Response Theory (lRT)? 2. Why is IRT useful in CAT? 3. What aspects of the item's performance are characterized with IRT? 4. How is the examinee's proficiency characterized? 5. What are some methods of estimating examinee proficiency? Describe each method's advantages and disadvantages? 6. There are many different IRT models. Compare and contrast the advantages and disadvantages of the I-PL, 2-PL, and 3-PL. 7. It has been pointed out that a prior is just another item. (a) Explain why this is true, and how it can be used to aid in the estimation of parameters. (b) Explain why this is false, and how it may cause unacceptable biases in estimates of individual students' proficiencies when scores are used in high-stakes decisions. 8. How do we measure how accurately an examinee's proficiency has been estimated? 9. Once a CAT is operational, new items have to be calibrated online. What difficulties are inherent in doing this? How can they be eased? 10. Why should items with high values of c be avoided? 11. Why do we seek items with high values of a? 12. What is the difference between the ideal distribution of bs in a CAT as opposed to that in a P&P test? Why? 13. What are the problems associated with conditional and joint maximumlikelihood estimation that are solved using marginalization? 14. What are the key assumptions underlying IRT? How sensitive is the validity of CAT to these assumptions? 15. How is the assumption of local independence threatened by traditionally structured paragraph comprehension items (items with a reading passage followed by a sequence of questions)? Is this threat important? How can this threat be ameliorated?

Angoff, W. H., & Huddleston, E. M. (1958). The multi-level experiment: A study of a two-stage system for the College Board Scholastic Aptitude Test. Statistical Report 58-21. Princeton, NJ: Educational Testing Service.

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. Psychometrika. 37. 29-51.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information factor analysis. Applied Psychological Measurement. 12. 261-280.

Bock, R. D., & Mislevy, R. J. (1981). The profile 0/ American youth: Data quality analysis o/the Armed Services Vocational Aptitude Battery. Chicago: National Opinion Research Center.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement. 6. 431-444.

Bock, R. D., & Mislevy, R. J. (1988). Comprehensive educational assessment for the states: The duplex design. Educational Evaluation and Policy Analysis. 10. 89-105.

Brown, J. M., & Weiss, D. J. (1977). An adaptive testing strategy for achievement test batteries. (Research Report No. 77-6). Minneapolis: University of Minnesota, Psychometric Methods Program.

Cronbach, L. J., & GIeser, G. C. (1965). Psychological tests and personnel decisions (2nd ed.). Urbana: University of Illinois Press.

De Groot, M. H. (1970). Optimal statistical decisions. New York: McGraw-Hili.

Greaud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. Applied Psychological Measurement. 10, 23-34.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement. 21. 347-360. .

Hansen, D. N. (1969). An investigation of computer-based

science testing. In R. C. Atkinson & H. A. Wilson (Eds.). Computer-assisted instruction: A book o/readings (pp. 209-226). New York: Academic Press.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow-Jones Irwin.

Jensema, C. J. (I 974a). The validity of Bayesian tailored testing. Educational and Psychological Measurement, 34. 757-766. 5. TESTING ALGORITHMS 131

Jensema, C. J. (I 974b). An application of latent trait mental test theory. British Journal of Mathe

matical and Statistical Psychology. 27. 29-48.

Jensema, C. J. (1977). Bayesian tailored testing and the influence of item bank characteristics.

Applied Psychological Measurement. I. 111-120.

Killcross, M. C. (1976). A review of research in tailored testing (Report APRE No. 9176). Fran

borough, Hants, England: Ministry of Defense, Army Personnel Research Establishment.

Krathwohl, D. R., & Huyser, R. J. (1956). The sequential item test (SIT). American Psychologist.

2.419.

Lewis, c., & Sheehan, K. (1988). Using Bayesian decision theory to design a Computerized

Mastery Test. Princeton, NJ: Educational Testing Service, Unpublished manuscript.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1969). The development and evaluation of several

programmed testing methods. Educational and Psychological Measurement. 29. 129-146.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.). Computer

assisted instruction. testing. and guidance (pp. 139-183) New York: Harper & Row.

Lord, F. M. (l97Ia). Robbins-Munro procedures for tailored testing. Educational and Psychologi

cal Measurement. 31. 3-31.

Lord, F. M. (\97Ib). The self-scoring flexilevel test. Journal of Educational Measurement. 8.

147-151.

Lord, F. M. (l97Ic). A theoretical study of two-stage testing. Psychometrika. 36. 227-242.

Lord, F. M. (1977). A broad-range test of verbal ability. Applied Psychological Measurement. I .

• 95-100.

Lord. F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale,

NJ: Lawrence Erlbaum Associates.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika. 47. 149-174.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military

setting. In D. J. Weiss (Ed.). New horizons in testing (pp. 223-236). New York: Academic

Press.

Mislevy, R. J., & Wu, P. K. (1988). Inferring examin"e ability when some item responses are

missing (Research Report 88-48-0NR). Princeton, NJ: Educational Testing Service.

Owen, R. J. (1969). A Bayesian approach to tailored testing (Research Report 69-92). Princeton,

NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adap

tive mental testing. Journal of the American Statistical

Association. 70. 351-356.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores.

Psychometric Monograph. No. 17. 34. Part 2.

Sympson, J. B., & Hetter. R. D. (1985, October). Controlling item-exposure rates in computerized

adaptive testing. Paper presented at the annual conference of the Military Testing Association,

San Diego.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika. 51,

567-577.

Thissen, D., Steinberg, I., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple

categorical-response models. Journal of Educational Measurement. 26, 247-260.

Urry, V. W. (1970). A monte carlo investigation of logistic test models. Unpublished doctoral

dissertation, Purdue University West Lafayette, IN.

Urry, V. W. (1977). Tailored testing: A successful application of item response theory. Journal of

Educational Measurement. 14, 181196.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for

testlets. Journal of Educational Measurement. 24. 185-201.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. Journal of Educa

tional Statistics. 12. 339-368.

Weiss, D. 1. (1974). Strategies of adaptive ability measurement. Research Report 74-5. Minneapolis: University

of Minnesota, Psychometric Methods Program.

Weiss, D. 1. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.

Wolfe, 1. H. (1985). Speeded tests-Can computers improve measurement? Proceedings of the 27th Annual Conference of the Military Testing Association (pp. 49-54). San Diego, CA: Naval Personnel Research & Development Center.

Zimowski, M. F. (1988, April). The duplex design: An evaluation of the two-stage testing procedure. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Zimowski, M. F., & Bock, R. D. (1987). Full-information factor analysis of test forms from the ASVAB CAT pool (MRC Report #87-1). Chicago: National Opinion Research Center.
5. TESTING ALGORITHMS 133 EXERCISES/STUDY QUESTIONS I. What are the three key issues in any testing algorithm? 2. What are the issues involved in choosing which item to begin a CAT with? 3. What are the issues governing what is the next item chosen? 4. How does a CAT know when to stop the test? 5. Are stopping rules different for tests with cut scores than with tests whose goal is accurate proficiency estimation? Why? How? 6. How does a testing algorithm use the Info Table to select items for presentation? 7. What is item exposure? Why need we concern ourselves with it? 8. Why is content balancing of tests a more difficult issue to resolve within a CAT than in a fixed format test? 9. How is "two-stage testing" a CAT? How is it different?

10. What is a "testletT' What problems in CAT does the concept of a testlet solve? This page intentionally left blank

6 Scaling and Equating

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), Test equating (pps. 55-69) New York: Academic Press.

Angoff, W. H. (1984). Scales, norms and equivalent scores. Princeton, NJ: Educational Testing Service.

Braun, H. I., & Holland, P. W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), Test equating (pp. 949) New York: Academic Press. 6. SCALING AND EQUATING 157

Cureton, E. E., & Tukey, J. W. (1951). Smoothing frequency distributions, equating tests and

preparing norms. American Psychologist, 6, 404. (Abstract).

Dorans, N. 1, & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test

editions. Applied Measurement in Education, 3, 245-254.

Dorans, N. 1, & Plake, B. S. (Eds.) (1990). Selecting samples for equating: To match or not to match

[special issue]. Applied Measurement in Education, 3(1).

Fairbank, B. A. (1987). The use of presmoothing and postsmoothing to increase the precision of

equipercentile equating. Applied Psychological Measurement. 11, 245-262.

Flanagan, 1. C. (1939). The Cooperative Achievement Tests: A bulletin reporting the basic princi

ples and procedures used in the development of their system of scaled scores. New York:

American Council on Education, Cooperative Test Service.

Flanagan, J. C. (1951). Units, scores and norms. In E. F. Lindquist (Ed.), Educational measure

ment (pp. 695-763) Washington, DC: American Council on Education.

Holland, P. W., & Thayer, D. T. (1989). The kernal method of equating score distributions.

(RR-89-84). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Rubin, D. B. (1982). Test equating. New York: Academic Press.

Hull, C. L. (1922). The conversion of test scores into series which shall have any assigned mean

and degree of dispersion. Journal of Applied Psychology, 6, 298-300.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores.

Psychometrika, 27, 59-72.

Kelley, T. L. (1947). Fundamentals of statistics. Cambridge: Harvard University Press.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. Journal of

Educational Statistics, 9, 25-44.

Lawrence, I. M., & Dorans, N. J. (1990). The effect on equating results of matching samples on an

anchor test. Applied Measurement in Education, 3, 19-36.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (in press). What combination of sampling and

equating works best? Applied Measurement in Education.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale,

NJ: Lawrence Erlbaum Associates.

McCall, W. A. (1939). Measurement. New York: Macmillan.

Pearson, K. (1913). On the relationship of intelligence to size and shape of head, and to other

physical and mental characteristics. Biometrika, 5. 105-146.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L.

Linn (Ed.), Educational Measurement, (3rd ed. (pp. 221-262». New York: Macmillan.

Petersen, N. S., Marco, G. L., & Steward, E. E. (1982). A test of the adequacy of linear score

equating models. In P. W. Holland & D. R. Rubin (Eds.), Test equating (pp. 71-135) New

York: Academic Press.

Potthoff, R. F. (1982). Some issues in test equating. In P. W. Holland & D. B. Rubin (Eds.). Test

equating (pp. 201-242) New York: Academic Press.

Reckase, M. D., Ackerman, T. A., Carlson, J. E. (1988). Building a unidimensional test using

multidimensional items. Journal 0" Educational Measurement, 25, 193-203.

Thorndike, E. L., Bregman, E. 0., Cobb, M. V., & Woodyard, E. (1926). The measurement of

intelligence. New York: Columbia University, Teachers College, Bureau of Publications.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. Journal of

Educational Psychology, 16,433-451.

Tukey, 1. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Vale, C. D. (1986). Linking item parameters onto a common scale. Applied Psychological Mea

surement, 10, 333-344.

difference between a score and a scale? 2. Why is "percent correct" not a useful score on a CAT? 3. What are three kinds of scoring schemes that can be sensibly used in CAT? 4. What are the advantages of a percentile derived scale? A normalized scale? 5. What does the term equating mean when applied to different test forms? To CAT? 6. What must be true for a procedure to be properly called an equating method? 7. What does the term construct mean with respect to testing? 8. Can we consider a linear prediction equation obtained by regression scores on test form X against those on test form Y an equating function? If not, why not? If so, justify. 9. If test form X is less reliable than test form Y, can they ever be equated? Hint: remember the equity condition and consider examinees with below mean scores versus those with above mean scores. Can such an equating satisfy the symmetry condition? lO. What are the advantages of equipercentile equating? The disadvantages? 11. What is the preferred kind of equating to use with CAT?

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: Author.

Anastasi, A. (1988). Psychological testing. New York: Macmillan.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories o/mental test scores (pp. 392-479). Reading, MA: Addison-Wesley.

Cronbach, L. J. (1984). Essentials o/psychological testing. New York: Harper & Row.

Dorans, N. J. (1984). Approximate IRT formula score and scaled score standard errors o/measurement at different ability levels (Statistical Repon SR-84-118). Princeton, NJ: Educational Testing Service.

Green, B. F., Bock, R. D., Humphreys. L. G .. Linn. R. L., & Reckase. M. D. (I 984a). Evaluation plan/or the computerized adaptive vocational aptitude battery (MPL TN 85-1). San Diego. CA: Manpower and Personnel Laboratory. NPRDC.

Green. B. F .• Bock. R. D .• Humphreys. L. G .• Linn. R. L.. & Reckase. M. D. (l984b). Technical guidelines for assessing computerized adaptive tests. Journal 0/ Educational Measurement. 21. 347-360.

Kelley. T. L. (1927). The interpretation 0/ educational measurements. New York: World Book.

Kuder. G. F .• & Richardson. M. W. (1937). The theory of the estimation of test reliability. Psychometrika. 2. 151-160.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. Educational and Psychological Measurement. 13. 517-548.

Lord, F. M., & Novick, M. R. (1968). Statistical theories 0/ mental test scores. Reading, MA: Addison-Wesley.

McBride, J. R., & Manin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), New horizons in testing (pp.

223-236). New York: Academic Press.

Oosterloo, S. (1984). Confidence intervals for test information and relative efficiency. Statistica Neerlandica. 38. 37-53.

Otis, A. S. (1925). Statistical method in educational measurement. New York: World Book.

Samejima, F. (1977). The use of the information function in tailored testing. Applied Psychological Measurement. I. 233-247.

184 THISSEN EXERCISES/STUDY QUESTIONS 1. What does the tenn reliability connote with respect to testing? 2. Why is the traditional measure of standard error inadequate on any test scored with IRT? 3. If reliability is outmoded on a CAT, what is it to be replaced by? 4. What are the advantages and disadvantages of scaling measurement error in the () metric? In the expected score metric? 5. How does the variance introduced by item sampling affect measurement precision? 6. How can we tell the difference between instability of test scores over time, and examinee growth/leaming?

8 Validity

American Psychological Association (1985). Standards for educational and psychological testing.

Washington, DC: Author.

Anastasi, A. (1988). Psychological testing. New York: Macmillan.

Bentler, P. M. (1985). Theory and implementation of EQS, a structural equations program. Los

Angeles: BMDP Statistical Software.

Bock, R. D. (1984). Full information factor analysis. Paper presented at the annual meeting of the

Psychometric Society, Santa Barbara, CA.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters:

An application of the EM algorithm. Psychometrika. 46, 443-449.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. Applied

Psychological Measurement. 12. 261-280.

Bond, L. (1981). Bias in mental tests. In B. F. Green (Ed.), Issues in testing: Coaching, disclosure,

and ethnic bias (pp. 55-77). San Francisco: Jossey-Bass.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi

method-multitrait matrix. Psychological Bulletin. 56. 81-105.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or

between tests. Psychometrika. /0. 1-19.

Cattell, J. McK. (1890). Mental tests and measurements.

Mind, 15, 373-381.

Cattell, J. McK., & Farrand, L. (1896). Physical and mental measurements of the students at Columbia University. Psychological Review, 3, 618-648.

Cronbach, L. J., GIeser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability ofbehavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Dorans, N. J., & Drasgow, F. (1978). Alternative weighting schemes for linear prediction. Organizational Behavior and Human Performance, 21, 316-345.

Dorans, N. J., & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23. 355-368.

Dorans, N. J., & Lawrence, I. (1987). The internal construct validity of the SAT. (Research Repon No. 87-35). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1988). The standardization approach to differential speededness. (Research Repon No. 88-31). Princeton, NJ: Educational Testing Service.

Dunbar, S. B., Mayekawa, S., & Novick, M. R. (1985). Simultaneous estimation of regression functions for Marine Corps technical training specialties (ONR Technical Repon 85-1). Iowa City: The University of Iowa, CADA Research Group.

Dunbar, S. B., & Novick, M. R. (1985). On predicting success in training for males andfemales: Marine Corps clerical specialties and ASVAB Forms 6 and 7 (ONR Technical Repon 85-2). Iowa City: The University of Iowa, CADA Research Group.

Embretson, S. E. (1983). Construct validity: Construct representation and nomothetic span. Psychological Bulletin. 93. 179-197.

Embretson, S. E. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 195-218). New York: Academic Press.

Gilben, J. A. (1894). Researches on the mental and physical development of school children. Studies Yale Lab .• 2. 40-100.

Green, B. F. (1988). Construct validity of computer-based tests. In H. Wainer & H. Braun (Eds.), Test validity (pp. 77-86). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F., Bock, R. D., Humphreys, L. G .• Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement. 21. 347-360.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Hayduk, L. A. (1987). Structural equation modeling with L1SREL. Baltimore, MD: Johns Hopkins University Press.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), Test validity (pp. 192-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Houston, W. M., & Novick, M. R. (1987). Race-based differential prediction in Air Force technical training programs. Journal of Educational Measurement. 24. 309-320.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin. 96. 72-98.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. M. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. E. Reynolds & R. T. Brown, (Eds.) Perspectives on bias in mental testing (pp. 41-99). New York: Plenum.

Joreskog, K. J. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology (Volume II, pp. I-56). San Francisco: W. F. Freeman.

Joreskog, K. J., & Sorbom, D. (1979). Advances infactor analysis and structural equation models. Cambridge, MA: Abt Books. 8, VALIDITY 227

Joreskog, K. J., & Sorbom, D. (1984). USREL VI: Analysis of

linear structural equation models by

maximum likelihood and least squares methods. Mooresville, IN: Scientific Software.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), Latent

trait and latent class models (pps. 263-275). New York: Plenum.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S.

A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen,

Measurement and Prediction (pp. 362-412). New York: Wiley.

Lindley, D. L., & Smith, A. F. M. (1972). Bayesian estimates for the linear model. Journal of the

Royal Statistical Society (Series B), 34, 1-41.

Linn, R. L. (1978). Single-group validity, differential validity and differential prediction. Journal

of Applied Psychology, 63, 507-512.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of

reading comprehension. Applied Psychological Measurement, 5, 159-173.

Loehlin, J. C. (1987). Latent variable models: An introduction to factor, path and structural

analysis. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Pooninga

(Ed.), Basic problems in cross-cultural research (pp. 19-29). Amsterdam: Swets & Zeitlinger.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale,

NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA:

Addison-Wesley.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. Journal

of Educational Statistics, 11, 3-31.

Molenaar, I. W., & Lewis, C. (1979). An improved model and computer program for Bayesian m

group regression (ONR Technical Repon 79-5). Iowa City: The University of Iowa, College of

Education.

Moreno, K., Segall, D.O., & Kieckhaefer, W. F. (1985). A validity study of the computerized

adaptive testing version of the Armed Services Vocational Aptitude Battery. Proceedings of the

27th annual meeting of the Military Testing Association (pp. 29-33). San Diego, CA: Navy

Personnel Research and Development Center.

Muthen, B. O. (1987). USCOMP user's guide. Mooresville, IN: Scientific Software.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of cenain test criteria for

purposes of statistical inference. Biometrika, 20A, 174-240 and 263-294.

Novick, M. R., & Ellis, D. D. Jr. (1977). Equal opponunity in educational and employment

selection. American Psychologist, 32, 306-320.

Novick, M. R., & Jackson, P. H. (1974). Funher cross-validation analysis of the Bayesian m-group

regression method. American Educational Research Journal, 11, 77-85.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction.

New York: Holt, Rinehan, Winston.

Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. Braun (Eds.), Test

validity (pp. 49-59). Hillsdale, NJ: Lawrence Erlbaum Associates.

Reckase, M. D. (1981). The formation of homogeneous item sets when guessing is afactor in item

response (Research Repon No. 81-5). Columbia: University of Missouri, Department of Educa

tional Psychology.

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. New York: Holt, Rinehart, &

Winston.

Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. Journal

of the American Statistical Association. 75. 801-816.

Schmitt, A. P., & Dorans, N. J. (1988). Differential itemfunctioningfor minority examinees on the

SAT. (Research Repon No. 88-32). Princeton, NJ: Educational Testing Service.

Sympson, J. B., Weiss, D. J., & Ree, M. J. (1983). A validity comparison of adaptive testing in a military technical training environment. (AFHRL-TR-81-40). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Sympson, J. B., Weiss, D. J., & Ree, M. J. (1984, April). Predictive validity of computerized adaptive testing in a military training environment. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Thissen, D. (1988). Multilog: User's guide. Mooresville,

IN: Scientific Software.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), Test validity (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Urry, V. W. (1981). Tailored testing, its theory and practice. Part II: Ability and item parameter estimation, multiple ability application, and allied procedures. (NPRDC TR81). San Diego, CA: Navy Personnel Research and Development Center.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, 83, 213-217.

Wilson, D. T., Wood, R., & Gibbons, R. (1984). TESTFACT: Test scoring, item statistics, and item factor analysis. Mooresville, IN: Scientific Software.

Wissler, C. (1901). The correlation of mental and physical tests. Psychological Monographs, 3, No. 16, 1-62.

Wright, D. (1987). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. In A. P. Schmitt & N. J. Dorans (Eds.), Differential item functioning on the Scholastic Aptitude Test (RM-87-1), (pp. 1-27), Princeton, NJ: Educational Testing Service.

Zimowski, M. F., & Bock, R. D. (1987). Full-information item factor analysis of test forms from the ASVAB CAT pool (MRC Report No. 87-1, Revised). Chicago, IL: Methodology Research Center, National Opinion Research Center. 8. VALIDITY 229 EXERCISES/STUDY QUESTIONS I. What does the tenn validity mean in testing? 2. How is the phrase "The test was validated" incorrect? 3. How is the measurement of the validity of test-based inferences different for a CAT than for a fixed fonnat test? 4. What is content validity? 5. What is construct validity? 6. Why is it important to study the multivariate structure of a test in assessing its validity?,' 7. How is the multitrait-multimethod approach helpful in assessing mode effects on validity? 8. Is a CAT, administered without constraining time limits, likely to increase or decrease validity? Why 9. What is predictive validity? Why is this the most important kind of validity for most applications of testing?

10. It is noted that a test's predictive validity diminishes as it becomes increasingly utilized. What might be the cause of this widely observed phenomenon? Explain.

II. If the validity of a test is seriously affected by selection effects, how can these be adjusted?

12. How does one collect predictive validity data on a new test (one which has yet to be administered)?

13. If a new test (say a CAT) has a precursor (say a paper and pencil version), explain how equating the CAT to the P&P version aids in obtaining preliminary estimates of the CAT's validity.

14. What kinds of events can affect validity? Which of these are unique to CAT? To P&P?

IS. What is dij7 How is it different from item bias?

16. What statistical procedures can be used to detect dij7

17. What are the advantages of the Mantel-Haenszel approach? The disadvantages?

18. What are the advantages of the IRT-LR approach? The disadvantages? This page intentionally left blank

psychometric measures of inductive reasoning. In S. E. Embretson (Ed.), Test design: Develop

ments in psychology and psychometrics (pp. 77-147). New York: Academic Press.

Christofferson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Cliff, N. (1987). Analyzing multivariate data. New York: Harcourt Brace Jovanovich.

Cooper, L. A. (1982, August). Strategies and spatial skill. Invited address at the American Psycho

logical Association convention, Washington, DC.

Cooper, L. A. (1983). Analogue representations of spatial objects and transformations. In O. J.

Braddick & A. C. Sleigh (Eds.), Physical and biological processing of images (pp. 231-264).

New York: Springer-Verlag.

Cronbach, L. J. (1984). Essentials of psychological testing (Fourth Edition.) New York: Harper &

Row.

Crone, C. R., Folk, V. G., & Green, B. F. (1988). The effect of item exposure control on informa

tion and measurement error in CAT. (Research Report 88-1). Baltimore, MD: Psychology

Department, The Johns Hopkins University.

Crowne, D., & Marlowe, D. (1964). The approval motive. New York: Wiley.

Damarin, F. (1970). A latent structure model for answering questions. Psychological Bulletin, 73,

23-40.

Drasgow, F., Levine, M. V., & Mclaughlin, M. E. (1987). Detecting inappropriate test scores

with optimal and practical appropriateness indices. Applied Psychological Measurement, II, 59

80.

Embretson, S. E. (1985a). Test design: Developments in psychology and psychometrics. New York:

Academic Press.

Embretson, S. E. (I 985b). Multicomponent latent trait models for test design. In S. E. Embretson

(Ed.), Test design: Developments in psychology and psychometrics (pp. 195-218). New York:

Academic Press.

Furneaux, W. D. (1961). Intellectual abilities and problem solving behavior. In H. J. Eysenck

(Ed.), The handbook of abnormal psychology (pp. 167-192). London: Pittman.

Greaud, V. A. (1987). Investigation of the unidimensionality assumption of item response theory.

Unpublished doctoral dissertation, Johns Hopkins University.

Greaud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of

speed tests. Applied Psychological Measurement, /0, 23-34.

Green, B. F. (1988). The construct validity of computer-based tests. In H. Wainer and H. Braun

(Eds.) Test validity (pp. 77-86). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F. (1981). A primer of testing. American Psychologist, 36, 1001-1011.

Green, B. F. (l983a). Adaptive testing by computer. In R. B. Ekstrom (Ed.), Measurement,

technology, and individuality in education: New directions for testing and measurement, No. 17

(pp. 5-12). San Francisco, CA: Jossey-Bass.

Green, B. F. (I 983b). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), Principals

of modern psychological measurement (pp. 69-80). Hillsdale, NJ: Lawrence Erlbaum Asso

ciates.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. B., & Reckase, M. D. (1984). Technical

guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21,

347-360.

Hambleton, R. K. (1986, February). Effects of item order and anxiety on test peiformance and

stress. Paper presented at the annual meeting of Division D, the American Educational Research

Association, Chicago.

Hardwicke, S. B., & Yoes, M. E. (1984). Attitudes and peiformance on computerized vs. paper

and-pencil tests. San Diego, CA: Rehab Group.

Harman, H. H. (1976). Modern factor analysis (3rd ed.) Chicago: University of Chicago Press.

Hetter, R. D., & Segall, D. O. (1986). Relative precision of paper-and-pencil and computerized

266 WAINER ET AL. adaptive tests. Proceedings of the 28th annual conference of the Military Testing Association (pp. 13-18). Mystic. CT: U.S. Coast Guard.

Hoijtink. H. (1988). A latent trait model for dichotomous choice data. Unpublished manuscript. The Netherlands: University of Groningen.

Holland. P. W .• & Rubin. D. B. (1982). Test equating. New York: Academic Press.

Hulin. C. L.. Drasgow. F .• & Komocar. J. (1982). Applications of item response theory to analysis of

attitude scale translations. Journal of Applied Psychology. 67. 818-825.

Hunt. E. G .• Frost. N .• & Lunneborg. C. L. (1973). Individual differences in cognition: A new approach to intelligence. In G. H. Bower (Ed.). The psychology of learning and motivation (Vol. 7. pp. 87-122). New York: Academic Press.

Kuhl. J. (1985). Volitional mediators of cognition-behavior consistency: Self-regulatory processes and action versus state orientation. In J. Kuhl & J. Beckmann (Eds.). Action control: From cognition to behavior (pp. 101-\28). Berlin: Springer-Verlag.

Kuhl. J. (1978). Situations-. reaktionsund personbezeogene Konsistenz des Leistungsmotivs bei der Messung mittels des Heckhausen-TAT. Archiv fur Psychologie. 52. 37-52.

Lazarsfeld. P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer. L. Guttman. E. A. Suchman. P. F. Lazarsfeld. S. A. Star. & J. A. Clausen. Measurement and Prediction (pp. 362-412). New York: Wiley.

Leary. L. F .• & Dorans. N. J. (1985). Implications for altering the context in which test items appear: An historical perspective on an immediate concern. Review of Educational Research. 55. 387-413.

Lewis. C. & Sheehan. K. (1988). Using Bayesian decision theory to design a computerized mastery test. Unpublished manuscript. Princeton. NJ: Educational Testing Service.

Lewis. C. (in preparation). Validity-based scoring. Manuscript in preparation. Princeton. NJ: Educational Testing Service.

Likert. R. (1932). A technique for the measurement of attitudes. Archives of Psychology. (Whole No. 140).

Lord. F. M .• & Novick. M. R. (1968). Statistical theories of mental test scores. Reading. MA: Addison-Wesley.

Lord. F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale. NJ: Lawrence Erlbaum Associates.

MacNicol. K. (1956). Effects of varying order of item difficulty in an unspeeded verbal test. Unpublished

manuscript. Princeton. NJ: Educational Testing Service.

Masters. G. N .• & Wright. B. D. (1984). The essential process in a family of measurement models. Psychometrika. 49. 529-544.

Masters. G. N. (1982). A Rasch model for partial credit scoring. Psychometrika. 47. 149174.

Masters. G. N. (1985). A comparison of latent-trait and latent-class analyses of Likert-type data. Psychometrika. 50. 69-82.

McKinley. R. L.. & Reckase. M. D. (l983a). An extension of the two-parameter logistic model to the multidimensional latent space. (Research Report ONR83-2). Iowa City: The American College Testing Program.

McKinley. R. L.. & Reckase. M. D. (19831,). An application of a multidimensional extension of the two-parameter latent trait model. (Research Report ONR83-3). Iowa City: The American College Testing Program.

Messick. S. & Jungeblut. A. (1981). Time and method in coaching for the SAT. Psychological Bulletin. 89. 191-216.

Mislevy. R. J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics. I I, 3-31.

Mollenkopf. W. G. (1950). An experimental study of the effects on item analysis data of changing item placement and test-time limit. Psychometrika. 15, 291-315. 9.
FUTURE CHALLENGES 267

Monk, 1. 1., & Stallings, W. M. (1970). Effect of item order on test scores. Journal of Educational

Research. 63. 463-465.

Moreno, K. E. (1987). Military applicant testing: Replacing paper-and-pencil with computerized

adaptive tests. Paper presented at the 1987 American Psychological Association Conference,

New York.

Moreno, K. E., Wetzel, C. D., McBride, 1. R., & Weiss, D.

1. (1984). Relationship between

corresponding Armed Services Vocational Aptitude Battery and computerized adaptive testing

subtests. Applied Psychological Measurement. 8. 155-163.

Mulaik, S. A. (1972). The foundations of factor analysis. New York: McGraw-Hili.

Muraki, E., & Engelhard, G. (1985). Full-information item factor analysis: Applications of EAP

scores. Applied Psychological Measurement. 9. 417-430.

Muthen, B. (1978). Contributions to factor aflalysis of dichotomous variables. Psychometrika. 43.

551-660.

Muthen, B. (1981). Factor analysis of dichotomous variables: American attitudes toward abortion.

In D. 1. 1ackson & E. F. Borgatta (Eds.), Factor analysis and measurement in sociological

research (pp. 201-214). Beverly Hills, CA: Sage.

Muthen, B. (1987). USCOMP: Analysis of linear structural relations with a comprehensive mea

surement model. Mooresville, Indiana: Scientific Software.

Muthen, B., & Lehman, 1. (1985). Multiple group IRT modeling: Applications to item bias analy

sis. Journal of Educational Statistics. 10. 133-142.

Pa<;hella, R. G. (1974). The interpretation of reaction time in information-processing research. In

B. H. Kantowitz (Ed.), Human iriformation processing (pp. 41-82). Hillsdale, N1: Lawrence

Erlbaum Associates.

Parsons, C. K., & Hulin, C. L. (1982). An empirical comparison of item response theory and

hierarchical factor analysis in applications to the

measurement of job satisfaction. Journal of

Applied Psychology. 67. 826-834.

Peterson, N. G. (Ed.) (1987, May). Development and Field Test of the Trial Batteryfor Project A

(Technical Report 739). Alexandria. VA: U.S. Army Research Institute for the Behavioral and

Social Sciences.

Peterson, R. C. (1931). A scale for measuring attitude toward capital punishment. Chicago: Uni

versity of Chicago Press.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen:

Denmarks Paedagogiske Institut.

Robinson, 1. P., & Shaver, P. R. (1973). Measures of Social Psychological Attitudes. Institute for

Social Research, University of Michigan.

Rosenbaum, P. R. (1988). A note on item bundles. Psychometrika. 53. 349-360.

Rotter. 1. B. (1966). Generalized expectancies for internal versus external control of reinforcement.

Psychological Monographs. 80. (Whole No. 609).

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores.

Psychometrika Monographs. (Whole No. 17).

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional

latent space. Psychometrika. 39. 111-121.

Samejima. F. (1979). A new family of models for the multiple-choice item. Research Report 79-4,

Knoxville, TN: Department of Psychology, University of Tennessee.

Sands, W. A. (1985). An overview of the CAT-ASVAB program. Proceedings of the 27th annual

meeting of the Military Testing Association (pp. 19-22). San Diego, CA: Navy Personnel

Research and Development Center.

Sands, W. A., & Gade, P. A. (1983). An application of computerized adaptive testing in army

recruiting. Journal of computer-based instruction. 10. 37-89.

Sax, G., & Carr, A. (1962). An investigation of response sets on altered parallel forms.

Educational and Psychological Measurement. 22. 371-376.

Schaeffer, N. C. (1988). An application of item response theory to the measurement of depression. Sociological Methodology, 18, 271-307.

Scheiblechner, H. (1979). Specifically stochastic latency mechanisms. Journal of Mathematical Psychology, 19, 18-38.

Segall, D. O. (1989). Specifying time·limits on computerized adaptive tests from censored and uncensored time distributions. (Draft technical report). San Diego, CA: Navy Personnel Research and Development Center.

Steinberg, L. (1986, June). Likert revisited: The measurement of attitudes. Presented at the annual meeting of the Psychometric Society, Toronto.

Steinberg, L. (1989, July). What do these items measure? (Really?): Some considerations arising from item analysis. Presented at the annual meeting of the Psychometric Society, Los Angeles.

Sternberg, R. J. (1981). Testing and cognitive psychology. American Psychologist. 36. 11811189.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference (pp.

82-98). Minneapolis: University of Minnesota.

Sympson, J. B. (1983, June). A new lRT model for calibrating multiple-choice items. Paper presented at the annual meeting of the Psychometric Society, Los Angeles.

Sympson, J. B. (1985, August). Alternative objectives in test equating: Different goals imply different scales. Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Sympson, J. B. (1986. August). Extracting informationfrom wrong answers in computerized adaptive testing. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Sympson, J. B., Weiss, D. J., & Ree, M. J. (1982). Predictive validity of conventional and adaptive tests in an Air Force training environment. (AFHRL-TR-81-40). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Tatsuoka, K., & Tatsuoka, M. (1979). A model for incorporating response time data in scoring achievement test. (CERL Report No. E-7). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing (pp. 179203). New York: Academic Press.

Thissen, D. (1988). MULTILOG user's guide. Mooresville, Indiana: Scientific Software.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. Psychometrika. 49. 501-519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika. 51. 567-577.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. Psychological Bulletin. 104. 385-395.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. Journal of Educational Measurement. 26. 161-176.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond

group mean differences: The concept of item bias. Psychological Bulletin. 99. 118-128.

Thissen, D., Steinberg. L., & Mooney. J. (1989). Trace lines for testlets: A use of multiplecategorical response models. Journal of Educational Measurement. 26. 247-260.

Thissen, D., Steinberg. L., Pyszczynski. T .• & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. Applied Psychological Measurement. 7. 211-226.

Thomas, T. J. (1989). Item presentation controls for computerized adaptive testing: Content bal9. FUTURE CHALLENGES 269

ancing vs. mini-CAT (Research Report 89-1). Baltimore, MD: Psychology Department, The

Johns Hopkins University.

Thurstone, L. L. (1928). Attitudes can be measured. American Journal of Sociology, 33, 529-554.

Towle, N. 1., & Merrill, P. F. (1975). Effects of anxiety type and item difficulty sequencing on

mathematics test performance. Journal of Educational Measurement, 12, 241-249.

Vernon, P. E. (1979). Intelligence: Heredity and environment. San Francisco: Freeman.

Vicino, F. L., & Hardwicke, S. B. (1984, March). An evaluation of the utility of large scale

computerized adaptive testing. Paper presented at the American Educational Research Associa

tion convention, New Orleans, LA.

Wainer, H. (1983). On item response theory and computerized adaptive tests: The coming technical

revolution in testing. Journal of College Admissions, 28, 9-16.

Wainer, H. (1989). The future of item analysis. Journal of Educational Measurement, 26, 191

208.

Wainer, H., & Braun, H. (1988). Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for

testlets. Journal of Educational Measurement, 24, 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. Journal oj Educational Meas

urement, 27, 1-14.

Wainer, H., & Messick, S. (1983). Principals of modern psychological measurement. Hillsdale,

NJ: Lawrence Erlbaum Associates.

Wald, A. (1947). Sequential analysis, New York: Wiley.

Wang, M. M. (1987, June). Measurement bias in the application of a unidimensional model to

multidimensional item response data. Paper presented at the ONR conference on Model-Based

Measurement, Columbia, SC.

White, P. O. (1973). Individual differences in speed, accuracy, and persistence. In H. J. Eysenck

(Ed.), The measurement of intelligence (pps. 246-260). Lancaster, England: Medical and Tech

nical Publishing.

Wilbur, E. R. (1986). Design and development of the ACAP test item data base. Proceedings of the

28th annual conference of the Military Testing Association (pp. 601-605). Mystic, CT: U.S.

Coast Guard.

Wilson, D., Wood, R., & Gibbons, R. D. (1984). TESTFACT: Test scoring, item statistics, and

item factor analysis. Mooresville, Indiana: Scientific

Software.

Wohlwill, J. (1963). The measurement of scalability for non-cumulative items. Educational and

Psychological Measurement, 23, 543-555.

Wolfe, 1. H., Alderton, D. L., Cory, C. H., & Larson, G. E. (1987, March). Reliability and

validity of new computerized ability tests. In H. Baker & Laabs, G. 1. (Eds.), Proceedings of the

Department of Defense/Educational Testing Service conference on Job Performance Measure

ment Technologies (pp. 369-382). San Diego, CA: Navy Personnel Research and Development

Center.

 EXERCISES/STUDY QUESTIONS I. If a test, like a CAT, is self-paced, why are any time constraints necessary at all? 2. In a CAT, response time to items can be recorded easily. How can this information be included to improve the test? 3. Cheating is a problem on all tests. How are the effects of cheating diminished in a CAT? 4. In fixed format tests examinees can "skip" an item that is too difficult, with the option of returning later should they get an insight. Why is this not possible in a CAT? Should this option be made available? How might it work? 5. If the IRT model utilized in building and scoring a CAT does not fit, how do we know? What are the consequences? What are our options? 6. Describe how more complex characterization of responses can be used in scoring a CAT. What do we gain? What are the costs? 7. What is a testlet? 8. What problems does a testlet-based CAT solve? 9. How is a CAT able to provide a fairer test to the entire test-taking population? to. What are the advantages of a CAT for the testing of the handicapped? II. Most people taking a CAT will find that they get about 60% of the items correct. Although this may feel good to those who are accustomed to failing most tests, others may find this discouraging. Discuss how this might be explained to ameliorate such concerns. 12. A traditional test, administered with nothing more complex than a #2 pencil and an answer sheet, is cheap, easy, and reliable. Justify the additional expense involved in CAT testing. 13. If major testing programs (like the SAT or the ACT) moved toward CAT, discuss the implications for the average

examinee. 14. Once tests are administered on a computer, describe some testing options that would be open that are currently not available.

10 Caveats, Pitfalls, and Unexpected Consequences of Implementing Large-Scale Computerized Testing

Boswell, J. (1791). Life of Samuel Johnson. London: Corry.

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). Case studies in computer adaptive test

design through simulation (ETS Research Report 93-56). Princeton, NJ: Educational Testing Service.

Feynman, R. P. (1986). "Appendix F: Personal Observations on the Reliability of the Shuttle," in vol

ume II, p. F5 of the Report of the Presidential Commission on the Space Shuttle Challenger Ac

cident. Washington, DC: US Government Printing Office.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ:

Lawrence Erlbaum Associates.

McLeod, L. (1998). Alternative methods for the detection of item preknowledge in computerized adap

tive testing. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). Computerized adaptive testing: From

inquiry to operation. Washington, DC: American Psychological Association.

Segall, D.O., Moreno, K. E. Bloxom, B. M., & Hetter, R. D. (1997). Psychometric procedures for ad

ministering CAT ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride, (Eds.), Computerized

adaptive testing: From inquiry to operation (pp. 131-140). Washington, DC: American Psycho

logical Association.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive

testing. Applied Psychological Measurement. 17,277-292.

Stocking, M. L., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive test

ing. Applied Psychological Measurement. 22,271-279.

Wainer, H. et al. (1990). Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Lawrence Erlbaum

Associates.

Wainer, H. (1997). Rescuing Computerized Testing by Breaking Zipf's Law. Department of Statistics,

Cornell University, Ithaca, NY.

Wainer, H. (2000). Rescuing Computerized Testing by Breaking Zipf's Law. Journal of Educational

and Behavioral Statistics. 25.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. Educational Measurement: Issues and Practice. 17, 17-27.

Way, W. D., & Steffen M. (1998, April). Strategies for managing item pools to maximize item security. Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.

Way, W. D., Steffen, M., & Anderson, G. S. (1998, September). Developing. maintaining. and renewing the item inventory to support computer-based testing. Paper presented at colloquium "Computer-Based Testing: Building the Foundation for Future Assessments," Philadelphia.

Wise, L. L., Curran, L. T., & McBride, J. R. (1997). CAT-ASVAB cost and benefit analyses. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing: From inquiry to operation (pp. 227-238). Washington, DC: American Psychological Association.

Wolfe, J. H., Alderton, D. L., Cory, C. H., & Larson, G. E. (1987). Reliability and validity of new computerized ability tests. In H. Baker & G. J. Laabs (Eds.), Proceedings of the Department of Defense/Educational

Testing Service conference on Job Performance Measurement Technologies (pp. 369-382). San Diego, CA.: Navy Personnel Research and Development Center.

Wolfe, J. H., Alderton, D. L., Cory, C. H., Larson, G. E., Bloxom, B. M., & Wise, L. L. (1997). Expanding the content of the CAT-ASVAB: New tests and their validity. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.). Computerized adaptive testing: From inquiry to operation (pp. 239-249). Washington, DC: American Psychological Association.

Zipf, G. K. (1949). Human Behavior and the Principle of least effort. Cambridge, MA: Addison-Wesley. 10. IMPLEMENTING LARGE-SCALE COMPUTERIZED TESTING 299 EXERCISES/STUDY QUESTIONS

1. Explain why a computerized test must be administered continuously rather than on a limited number of test dates.

2. Give three reasons why computerizing a test will limit access (see chapter 2 for some additional hints).

3. What is Zipf's Law and why does it have a direct bearing on test security for CATs?

4. Discuss the role that simulation plays in building CAT item pools.

5. How did Kaplan steal the GRE item pool?

6. What steps can be taken that could keep the Kaplan strategy from working again?

7. Why would low-scoring examinees benefit more from stolen items from sources of modest ability than from sources of high ability? This page intentionally left blank

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*, 251–269.

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests.* Washington, DC: American Psychological Association.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitudes. *Applied Psychological Measurement, 12*, 33–51.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, DC: American Council on Education.

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 55–69). New York: Academic Press.

Angoff, W. H. (1984). *Scales, norms and equivalent scores.* Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Dyer, H. S. (1971). The Admissions Testing Program. In W. H. Angoff (Ed.), *The College Board Admissions Testing Program* (pp. 1–13). New York: College Entrance Examination Board.

Angoff, W. H., & Huddleston, E. M. (1958). *The multi-level experiment: A study of a two-stage system for the College Board Scholastic Aptitude Test* (Statistical Report 58-21). Princeton, NJ: Educational Testing Service.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17*, 283–296.

Bejar, I. I. (1985). Speculations on the future of test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 279–294). New York: Academic Press.

Bentler, P. M. (1985). *Theory and implementation of EQS. a structural equations program*. Los Angeles: BMDP Statistical Software.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392–479). Reading, MA: Addison-Wesley.

Bloxom, B. (1985). Considerations in psychometric modeling of response time. *Psychometrika, 50*, 383–397.

Bloxom, B. (1992). *Accelerated CAT-AS VAB program: Psychometric decisions list* (Technical Report). San Diego, CA: Navy Personnel Research and Development Center.

Bloxom, B., & Vale, C. D. (1987, June). *Adaptive estimation of a multidimensional latent trait*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37*, 29–51.

Bock, R. D. (1984, June). *Full information factor analysis*. Paper presented at the annual meeting of the Psychometric Society, Santa Barbara, CA.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443–449.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12*, 261–280.

Bock, R. D., & Mislevy, R. J. (1981). *The profile of American youth: Data quality analysis of the Armed Services Vocational Aptitude Battery*. Chicago: National Opinion Research Center.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431–444.

Bock, R. D., & Mislevy, R. J. (1988). Comprehensive educational assessment for the states: The duplex design. *Educational Evaluation and Policy Analysis, 10*, 89–102.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275–285.

Bond, L. (1981). Bias in mental tests. In B. F. Green (Ed.), *Issues in testing: Coaching, disclosure, and ethnic bias* (pp. 55–77). San Francisco: Jossey-Bass.

Boswell, J. (1791). *Life of Samuel Johnson*. London: Corry.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

Braun, H. I., & Holland, P. W. (1982). *Observed score test equating: A mathematical analysis of some ETS equating procedures*. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Psychometric Methods Program.

Burt, C. (1911). Experimental tests of higher mental processes and their relation to general intelligence. *Journal of Experimental Pedagogy, 1*, 93–112.

Butterfield, E. C., Nielsen, D., Tangen, K. L., & Richardson, M. B. (1985). *Theoretically based psychometric measures of inductive reasoning*. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 77–147). New York: Academic Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multimethod-multitrait matrix. *Psychological Bulletin, 56*, 81–105.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika, 10*, 1–19.

Cattell, J. McK. (1890). Mental tests and measurements. *Mind, 15*, 373–381.

Cattell, J. McK., & Farrand, L. (1896). Physical and mental measurements of the students at Columbia University. *Psychological Review, 3*, 618–648.

Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*, 5–32.

Cliff, N. (1987). *Analyzing multivariate data*. New York: Harcourt Brace.

College Entrance Examination Board. (1981). *An SAT: Test and technical data for the Scholastic Aptitude Test administered in April 1981*. New York: Author.

Cooper, L. A. (1982, August) *Strategies and spatial skill*. Invited address at the meeting of the American Psychological Association, Washington, DC.

Cooper, L. A. (1983). Analogue representations of spatial objects and transformations. In O. J. Braddick & A. C. Sleigh (Eds.), *Physical and Biological Processing of Images* (pp. 231–264). New York: Springer-Verlag.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist. 12*, 671–684

Cronbach, L. J. (1984). *Essentials of psychological testing*. New York: Harper & Row.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Crone, C. R., Folk, V. G., & Green, B. F. (1988). *The effect of item exposure control on information and measurement error in CAT* (Research Report 88-1). Baltimore, MD: The Johns Hopkins University, Psychology Department.

Crowne, D., & Marlowe, D. (1964). *The approval motive*. New York: Wiley.

Cureton, E. E., & Tukey, J. W. (1951). Smoothing frequency distributions, equating tests and preparing norms. *American Psychologist, 6*, 404. (Abstract).

Damarin, F. (1970). A latent structure model for answering personal questions. *Psychological Bulletin, 73*, 23–40.

De Groot, M. H. (1970). *Optimal statistical decisions*. New York: McGraw-Hill.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, *39*, 1–38.

Department of Defense (1986). *A review of the development and implementation of the ASVAB forms 11, 12, & 13*. Washington, DC: Author.

Divgi, D. R., & Stoloff, P. H. (1986). *Effect of the medium of administration on ASVAB item response curves* (CNA 86-24). Alexandria, VA: Center for Naval Analysis.

Dorans, N. J. (1984). *Approximate IRT formula score and scaled score standard errors of measurement at different ability levels* (Statistical Report SR-84-118). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Drasgow, F. (1978). Alternative weighting schemes for linear prediction. *Organizational Behavior and Human Performance, 21*, 316–345.

Dorans, N. J., & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368.

Dorans, N. J., & Lawrence, I. M. (1987). *The internal construct validity of the SAT* (Research Report No. 87-35). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education, 3*, 245–254.

Dorans, N. J., & Plake, B. S. (Eds.). (1990). Selecting samples for equating: To match or not to match [Special issue]. *Applied Measurement in Education, 3(1)*.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1988). *The standardization approach to differential speededness* (Research Report No. 88-31). Princeton, NJ: Educational Testing Service.

Downey, M. T. (1965). *Ben T. Wood, educational reformer*. Princeton, NJ: Educational Testing Service.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59–80.

DuBois, P. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon.

Dunbar, S. B., & Novick, M. R. (1985). *On predicting success in training for males and females: Marine Corps clerical specialities and ASVAB Forms 6 and 7* (ONR Technical Report 85-2). Iowa City: The University of Iowa, CADA Research Group.

Dunbar, S. B., Mayekawa, S. & Novick, M. R. (1985). *Simultaneous estimation of regression functions for Marine Corps technical training specialities* (ONR Technical Report 85-1). Iowa City: The University of Iowa, CADA Research Group.

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*, 599–635.

Edgeworth, F. Y. (1892). Correlated Averages. *Philosophical Magazine* (5th series), *34*, 190–204.

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation* (ETS Research Report 93-56). Princeton, NJ: Educational Testing Service.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Embretson, S. E. (1983). Construct validity: Construct representation and nomothetic span. *Psychological Bulletin, 93*, 179–197.

Embretson, S. E. (1985a). *Test design: Developments in psychology and psychometrics*. New York: Academic Press.

Embretson, S. E. (1985b). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 195–218). New York: Academic Press.

Fairbank, B. A. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement, 11*, 245–262.

Feynman, R. P. (1986). "Appendix F: Personal Observations on the Reliability of the Shuttle," in volume II, p. F5 of the *Report of the Presidential Commission on the Space Shuttle Challenger Accident*. Washington, DC: US Government Printing Office.

Flanagan, J. C. (1939). *The Cooperative Achievement Tests: A bulletin reporting the basic principles and procedures used in the development of their system of scaled scores*. New York: American Council on Education, Cooperative Test Service.

Flanagan, J. C. (1951a). The use of comprehensive rationales in test development, *Educational and Psychological Measurement, 11*, 151–155.

Flanagan, J. C. (1951b). Units, scores and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). Washington, DC: American Council on Education.

Furneaux, W. D. (1961). Intellectual abilities and problem solving behavior. In H. J. Eysenck (Ed.), *The handbook of abnormal psychology* (pp. 167–192). London: Pittman.

Gibbons, R. D., Bock, R. D., & Hedeker, D. (1987, June). *Approximating multivariate normal orthant probabilities using the Clark algorithm*. Paper presented at the annual meeting of the Psychometric Society, Montreal.

Gilbert, J. A. (1894). Researches on the mental and physical development of school children. *Studies Yale Lab, 2*, 40–100.

Goodenough, F. L. (1926). *Measurement of intelligence by drawings*, Yonkers, NY: World Book.

Gould, J. D., Alfaro, L., Finn, R., Haupt, B., & Minuto, A. (1987). Reading from CRT displays can be as fast as reading from paper. *Human Factors, 29*, 497–517.

Greaud, V. A. (1987, April). *Investigation of the unidimensionality assumption of item response theory*. Unpublished doctoral dissertation, Johns Hopkins University, Baltimore, MD.

Greaud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10*, 23–34.

Green, B. F. (1981). A primer of testing. *American Psychologist, 36*(10), 1001–1011.

Green, B. F. (1983a). Adaptive Testing by Computer. In R. B. Ekstrom (Ed.), *Measurement, technology, and individuality in education: New directions for testing and measurement*, No. 17 (pp. 5–12). San Francisco, CA: Jossey-Bass.

Green, B. F. (1983b). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 69–80). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F. (1988). Construct validity of computer-based tests. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 77–86). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984a). *Evaluation plan for the computerized adaptive vocational aptitude battery* (MPL TN 85-1) San Diego, CA: Manpower and Personnel Laboratory, NPRDC.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. B., & Reckase, M. D. (1984b). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347–360.

Green, B. F., Bock, R. D., Linn, R. L., Lord, F. M., & Reckase, M. D. (1983). *A plan for scaling the Computerized Adaptive ASVAB*. Baltimore, MD: Johns Hopkins University, Department of Psychology.

Gulliksen, H. O. (1987). *A theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum Associates. (Original work published in 1950. New York: Wiley)

Hambleton, R. K. (1986, February). *Effects of item order and anxiety on test performance and stress*. Paper presented at the annual meeting of Division D, the American Educational Research Association, Chicago.

Hansen, D. N. (1969). An investigation of computer-based science testing. In R. C. Atkinson & H. A. Wilson (Eds.), *Computer-assisted instruction: A book of readings* (pp. 209–226). New York: Academic Press.

Hardwicke, S. B., Cooper, R., Eastman, L., & Vicino, F. L. (1984). *Computerized adaptive testing: A user manual*. San Diego, CA: Navy Personnel Research and Development Center.

Hardwicke, S. B., & White, K. D. (1983). *Predictive utility evaluation of computerized adaptive testing: Results of the Navy research*. San Diego, CA: Rehab Group.

Hardwicke, S. B., & Yoes, M. E. (1984). *Attitudes and performance on computerized vs. paper-and-pencil tests*. San Diego, CA: Rehab Group.

Harmon, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL*. Baltimore, MD: Johns Hopkins University Press.

Hetter, R. D., & Segall, D. O. (1986). Relative precision of paper-and-pencil and computerized adaptive tests. *Proceedings of the 28th annual conference of the military testing association* (pp. 13–18). Mystic, CT: U. S. Coast Guard.

Hoijtink, H. (1988) *A latent trait model for dichotomous choice data*. Unpublished manuscript. University of Groningen, The Netherlands.

Holland, P. W. & Rubin, D. B. (1982). *Test equating*. New York: Academic Press.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Thayer, D. T. (1989). *The kernal method of equating score distributions* (RR-89-84). Princeton, NJ: Educational Testing Service.

Houston, W. M., & Novick, M. R. (1987). Race-based differential prediction in Air Force technical training programs. *Journal of Educational Measurement, 24*, 309–320.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67*, 818–825.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow-Jones Irwin.

Hull, C. L. (1922). The conversion of test scores into series which shall have any assigned mean and degree of dispersion. *Journal of Applied Psychology, 6*, 298–300.

Humphreys, L. G. (1976). A factor model for research on intelligence and problem solving. In L. Resnick, (Ed.), *The nature of intelligence* (pp. 329–339). New York: Wiley.

Hunt, E. G., Frost, N., & Lunneborg, C. L. (1973). Individual differences in cognition: A new approach to intelligence. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 7, pp. 87–122). New York: Academic Press.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. M. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. E. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 41–99). New York: Plenum.

Hunter, R. V., & Slaughter, C. D. (1980). *ETS test sensitivity review process*. Princeton, NJ: Educational Testing Service.

James-Jones, G. (1986). Design and development of the ACAP test administration software. *Proceedings of the 28th annual conference of the military testing association* (pp. 612–617). Mystic, CT: U. S. Coast Guard.

Jenckes, T. A. (1868). *Civil Service of the United States* (Report No. 47, 40th Congress, 2nd Session, May 25).

Jensema, C. J. (1974a). The validity of Bayesian tailored testing. *Educational and Psychological Measurement, 34*, 757–766.

Jensema, C. J. (1974b). An application of latent trait mental test theory. *British Journal of Mathematical and Statistical Psychology, 27*, 29–48.

Jensema, C. J. (1977). Bayesian tailored testing and the influence of item bank characteristics. *Applied Psychological Measurement, 1*, 111–120.

Jöreskog, K. J. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Volume II, pp. 1–56). San Francisco: Freeman.

Jöreskog, K. J., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Jöreskog, K. J., & Sörbom, D. (1984). *LISREL VI: Analysis of linearstructural equation models by maximum likelihood and least squares methods*. Mooresville, IN: Scientific Software.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27*, 59–72.

Kelley, T. L. (1927). *The interpretation of educational measurements*. New York: World Book.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge: Harvard University Press.

Killcross, M. C. (1976). A review of research in tailored testing (Report APRE No. 9/76). Franborough, Hants, England: Ministry of Defense, Army Personnel Research Establishment.

Kingston, N. (1987). *Feasibility of using IRT-Based methods for Divisions D, E and I of the Architect Registration Examination* (Report prepared for the National Council of Architectural Registration Boards). Princeton, NJ: Educational Testing Service.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9*, 281–288.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 263–275). NewYork: Plenum.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9*, 25–44.

Krathwohl, D. R., & Huyser, R. J. (1956). The sequential item test (SIT). *American Psychologist, 2*, 419.

Kreitzberg, C. B., & Jones D. H. (1980). *An empirical study of the broad-range tailored test of verbal ability* (RR-80-5). Princeton, NJ: Educational Testing Service.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Kuhl, J. (1978). Situations-, reaktions- und personbezeogene Konsistenz des Leistungsmotivs bei der Messung mittels des Heckhausen-TAT. *Archiv fur Psychologie, 52*, 37–52.

Kuhl, J. (1985). Volitional mediators of cognition-behavior consistency: Self-regulatory processes and action versus state orientation. In J. Kuhl & J. Beckmann (Eds.), *Action control: From cognition to behavior* (pp. 101–128). Berlin: Springer-Verlag.

Lawrence, I. M., & Dorans, N. J. (1990). The effect on equating results of matching on an anchor test. *Applied Measurement in Education, 3*, 19–36.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). New York: Wiley.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: An historical perspective on an immediate concern. *Review of Educational Research, 55*, 387–413.

Levine, M. (1985). The trait in latent trait theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 41–65). Minneapolis, MN: University of Minnesota, Computerized Adaptive Testing Laboratory, Department of Psychology.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367–386.

Lewis, C. *Validity-based scoring*. Manuscript in preparation.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, (Whole No. 140).

Lindley, D. L., & Smith, A. F. M. (1972). Bayesian estimates for the linear model. *Journal of the Royal Statistical Society* (Series B), *34*, 1–41.

Link, H. C. (1919). *Employment psychology*. New York: Macmillan.

Linn, R. L. (1978). Single-group validity, differential validity and differential prediction. *Journal of Applied Psychology, 63*, 507–512.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 159–173.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement, 29*, 129–146.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating works best? *Applied Measurement in Education, 3*, 73–96.

Loehlin, J. C. (1987). *Latent variable models: An introduction to factor, path and structural analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability, *Psychological Monographs, 61*, No. 4.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517–548.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp. 139–183). New York: Harper & Row.

Lord, F. M. (1971a). Robbins-Munro procedures for tailored testing. *Educational and Psychological Measurement, 31*, 3–31.

Lord, F. M. (1971b). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*, 147–151.

Lord, F. M. (1971c). A theoretical study of two-stage testing. *Psychometrika, 36*, 227–242.

Lord, F. M. (1977a). A broad-range test of verbal ability. *Applied Psychological Measurement, 1*, 95–100.

Lord, F. M. (1977b). A study of item bias using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural research* (pp. 19–29). Amsterdam: Swets & Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacNicol, K. (1956). *Effects of varying order of item difficulty in an unspeeded verbal test*. Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Masters, G. N. (1985). A comparison of latent-trait and latent-class analyses of Likert-type data. *Psychometrika, 50*, 69–82.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49*, 529–544.

McBride, J. R. (1988, March) *A computerized adaptive version of the Psychological Corporation's Differential Aptitude Battery*. Paper presented at the annual meeting of the American Psychological Association, Atlanta, GA.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 223–236). New York: Academic Press.

McBride, J. R., & Sympson, J. B. (1985). The computerized adaptive testing system development project. In D. J. Weiss (Ed.), *Proceedings of the 1982 item response theory and computerized adaptive testing conference* (pp. 342–349). Minneapolis: University of Minnesota, Department of Psychology.

McCall, W. A. (1939). *Measurement*. New York: Macmillan.

McGucken, W. J. (1932). *The Jesuits and education*. Milwaukee: Bruce.

McKinley, R. L., & Reckase, M. D. (1983a). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Report ONR83-2). Iowa City: The American College Testing Program.

McKinley, R. L., & Reckase, M. D. (1983b). *An application of a multidimensional extension of the two-parameter latent trait model* (Research Report ONR83-3). Iowa City: The American College Testing Program.

McLeod, L. (1998). *Alternative methods for the detection of item preknowledge in computerized adaptive testing*. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*, 191–216.

Mislevy, R. J., & Stocking, M. L. (1987). *A consumer's guide to LOGIST and BILOG* (ETS Research Report 87-43). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381.

Mislevy, R. J. (1986a). Bayes modal estimation in item response models. *Psychometrika, 51*, 177–195.

Mislevy, R. J. (1986b). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3–31.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Bock, R. D. (1993). *BILOG 3. 04: Multiple-group IRT Analysis and Test Maintenance for Binary Items* [computer program]. Chicago, IL: Scientific Software, Inc.

Mislevy, R. J., Bock, R. D., & Muraki, E. (1988). *BIMAIN* [computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. S. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30*, 55–78

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57–75.

Mislevy, R. J., & Wu, P. K. (1988). *Inferring examinee ability when some item responses are missing* (Research Report 88-48-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Wingersky, M. S., & Kingston, M. (1990). *Evaluation of a procedure for calibrating "seeded" test items* (Final report to Battelle Columbus Division, Contract No. DAALO3-86-D-0001, Delivery Order 0708, Scientific Services Program). Princeton, NJ: Educational Testing Service.

Molenaar, I. W., & Lewis, C. (1979). *An improved model and computer program for Bayesian m-group regression* (ONR Technical Report 79-5). Iowa City: The University of Iowa, College of Education.

Mollenkopf, W. G. (1950). An experimental study of the effects on item analysis data of changing item placement and test-time limit. *Psychometrika, 15*, 291–315.

Monk, J. J., & Stallings, W. M. (1970). Effect of item order on test scores. *Journal of Educational Research, 63*, 463–465.

Moreno, K. E. (1987). *Military applicant testing: Replacing paper-and-pencil with computerized adaptive tests*. Paper presented at the meeting of the American Psychological Association, New York.

Moreno, K. E., Segall, D. O., & Kieckhaefer, W. F. (1985). A validity study of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 29–33). San Diego, CA: Navy Personnel Research and Development Center.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery and computerized adaptive testing subtests. *Applied Psychological Measurement, 8*, 155–163.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Muraki, E., & Engelhard, G. (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9*, 417–430.

Muthén, B. (1978). Contributions fo factor analysis of dichotomous variables. *Psychometrika, 43*, 551–660.

Muthén, B. (1981). Factor analysis of dichotomous variables: American attitudes toward abortion. In D. J. Jackson & E. F. Borgatta (Eds.), *Factor analysis and measurement in sociological research* (pp. 201–214). Beverly Hills, CA: Sage.

Muthén, B. (1987). *LISCOMP: Analysis of linear structural relations with a comprehensive measurement model*. Mooresville, IN: Scientific Software.

Muthén, B. & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*, 133–142.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20A*, 174–240, 263–294.

Novick, M. R., & Ellis, D. D. Jr. (1977). Equal opportunity in educational and employment selection. *American Psychologist, 32*, 306–320.

Novick, M. R., & Jackson, P. H. (1974). Further cross-validation analysis of the Bayesian m-group regression method. *American Educational Research Journal, 11*, 77–85.

Oosterloo, S. (1984). Confidence intervals for test information and relative efficiency. *Statistica Neerlandica, 38*, 37–53.

Otis, A. S. (1925). *Statistical method in educational measurement*. New York: World Book.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351–356.

Pachella, R. G. (1974). The interpretation of reaction time in information processing research. In B. H. Kantowitz (Ed.), *Human information processing* (pp. 41–82). Hillsdale, NJ: Lawrence Erlbaum Associates.

Parsons, C. K., & Hulin, C. L. (1982). An empirical comparison of item response theory and hierarchical factor analysis in applications to the measurement of job satisfaction. *Journal of Applied Psychology, 67*, 826–834.

Pearson, K. (1913). On the relationship of intelligence to size and shape of head, and to other physical and mental characteristics. *Biometrika, 5*, 105–146.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction*. New York: Holt, Rinehart & Winston.

Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 49–59). Hillsdale, NJ: Lawrence Erlbaum Associates.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. R. Rubin (Eds.), *Test equating*. New York, NY: Academic Press.

Peterson, N. G. (Ed.). (1987, May). *Development and Field Test of the Trial Battery for Project A* (Technical Report 739). Alexandria VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Peterson, R. C. (1931). *A scale for measuring attitude towards capital punishment*. Chicago: University of Chicago Press.

Porteus, S. D. (1915). Mental tests for the feebleminded: A new series, *Journal of Psycho-Asthenics, 19*, 200–213.

Potthoff, R. F. (1982). Some issues in test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 201–242). New York: Academic Press.

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). *Armed service vocational aptitude battery: Development of an adaptive item pool*. (Technical report 85-19). San Antonio, TX: Air Force Human Resources Laboratory.

Quan, B., Park, T. A., Sandahl, G., & Wolfe, J. H. (1984). *Microcomputer network for computerized adaptive testing (CAT)* (Technical report 84-33). San Diego, CA: Navy Personnel Research and Development Center.

Rafacz, B. A. (1986). Development of the test administrator's station in support of ACAP. *Proceedings of the 28th annual conference of the military testing association*. (pp. 606–611). Mystic, CT: U. S. Coast Guard

Rafacz, B. A. (1988). *A test administrator's user's manual: Developed in support of the accelerated CAT-ASVAB project (ACAP*; Revised). San Diego, CA: Navy Personnel Research and Development Center.

Rafacz, B. A., & Hetter, R. D. (1997). ACAP hardware selection, software development and acceptance testing. In W. A. Sands, B. Waters, & J. R. McBride (Eds.). *Computerized adaptive testing: From inquiry to operation* (pp. 145–156). Washington, DC: American Psychological Association.

Rafacz, B. A., & Moreno, K. E. (1987). *Functional requirements for the accelerated CAT-ASVAB project (ACAP)*. San Diego, CA: Navy Personnel Research and Development Center.

Rafacz, B. A., & Tiggle, R. B (1985). *Interactive screen dialogues for the examinee testing (ET) station developed in support of the accelerated CAT-ASVAB project (ACAP)*. San Diego, CA: Navy Personnel Research and Development Center.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Original work published in 1960. Copenhagen: Denmarks Paedagogiske Institut.)

Reckase, M. D. (1981). *The formation of homogeneous item sets when guessing is a factor in item response* (Research Report No. 81-5). Columbia: University of Missouri, Department of Educational Psychology.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*, 193–203.

Reed, A. V. (1979). Microcomputer display timing: Problems and solutions. *Behavior Research Methods and Instrumentation, 11*, 572–576.

Robinson, J. P., & Shaver, P. R. (Eds.). (1973). *Measures of social psychological attitudes*. Ann Arbor: Institute for Social Research, University of Michigan.

Rosenbaum, P. R. (1988). A note on item bundles. *Psychometrika, 53*, 349–360.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart, & Winston.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80*, (Whole No. 609)

Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association, 75*, 801–816.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 34* (17, Pt. 2).

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39*, 111–121.

Samejima, F. (1977). The use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 233–247.

Samejima, F. (1979). *A new family of models for the multiple choice item* (Research Report #79-4). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1983). Some methods and approaches of estimating the operating characteristics of discrete item responses. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 159–182). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sands, W. A. (1985). An overview of the CAT-ASVAB program. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 19–22). San Diego, CA: Navy Personnel Research and Development Center.

Sands, W. A., & Gade, P. A. (1983). An application of computerized adaptive testing in Army recruiting. *Journal of Computer-Based Instruction. 10*, 37–89.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Sax, G., & Carr, A. (1962). An investigation of response sets on altered parallel forms. *Educational and Psychological Measurement, 22*, 371–376.

Schaeffer, N. C. (1988). An application of item response theory to the measurement of depression. *Sociological Methodology, 18*, 271–307.

Scheiblechner, H. (1979). Specifically objective stochastic latency mechanisms. *Journal of Mathematical Psychology, 19*, 18–38.

Schmitt, A. P., & Dorans, N. J. (1987, August). Differential item functioning for minority examinees on the SAT. Paper presented at the annual meeting of the American Psychological Association, New York.

Schmitt, A. P., & Dorans, N. J. (1988). Differential item functioning for minority examinees on the SAT (Research Report No. 88-32). Princeton, NJ: Educational Testing Service.

Schratz, M. K. (1985). Assessment of the unidimensionality of the CAT-ASVAB subtests. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 34–37). San Diego, CA: Navy Personnel Research and Development Center.

Segall, D. O. (1987). *ACAP item pools: Analysis and recommendations* (Draft technical report). San Diego, CA: Navy Personnel Research and Development Center.

Segall, D. O. (1989). *Specifying time-limits on computerized adaptive tests from censored and uncensored time distributions*. (Draft technical report). San Diego, CA: Navy Personnel Research and Development Center.

Segall, D. O., & Moreno, K. E. (1986, March). *Dimensionality of the ACAP item pools: Findings and recommendations*. Paper presented at a meeting of the CAT-ASVAB technical committee, San Diego, CA.

Segall, D. O., Moreno, K. E. Bloxom, B. M., & Hetter, R. D. (1997). Psychometric procedures for administering CAT ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 131–140). Washington, DC: American Psychological Association.

Steinberg, L. (1986, June). *Likert revisited: The measurement of attitudes*. Paper presented at the annual meeting of the Psychometric Society, Toronto.

Steinberg, L. (1989, July). *What do these items measure? (Really?): Some considerations arising from item analysis*. Paper presented at the annual meeting of the Psychometric Society, Los Angeles.

Sternberg, R. J. (1981). Testing and cognitive psychology. *American Psychologist, 36*, 1181–1189.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277–292.

Stocking, M. L., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive testing. *Applied Psychological Measurement, 22*, 271–279.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 13–30). New York: Academic Press.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82–98). Minneapolis: University of Minnesota.

Sympson, J. B. (1983, June). *A new IRT model for calibrating multiple choice items*. Paper presented at the annual meeting of the Psychometric Society, Los Angeles.

Sympson, J. B. (1985, August). *Alternative objectives in test equating: Different goals imply different scales*. Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Sympson, J. B. (1986, August). *Extracting information from wrong answers in computerized adaptive testing*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego CA: Navy Personnel Research and Development Center.

Sympson, J. B., Weiss, D. J., & Ree, M. J. (1982). *Predictive validity of conventional and adaptive tests in an Air Force training environment* (AFHRL-TR-81-40). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Sympson, J. B., Weiss, D. J., & Ree, M. J. (1984, April). *Predictive validity of computerized adaptive testing in a military training environment*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Tatsuoka, K., & Tatsuoka, M. (1979). *A model for incorporating response time data in scoring achievement tests* (CERL Report No. E-7). Urbana, IL: University of Illinois, Cpmputer-based Education Research Laboratory.

Têng, Ssu-yü (1943). Chinese influence on the western examination system. *Harvard Journal of Asiatic Studies, 7*, 267–312.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203). New York: Academic Press.

Thissen, D. (1986). *Multilog: A user's guide*. Mooresville, IN: Scientific Software.

Thissen, D. (1988). *MULTILOG user's guide* (2nd ed.). Mooresville, IN: Scientific Software.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49*, 501–519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567–577.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*, 385–395.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26*, 161–176.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118–128.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247–260.

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement, 7*, 211–226.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thomas, T. J. (1989). *Item presentation controls for computerized adaptive testing: Content balancing vs. mini-CAT* (Research Report 89-1). Baltimore, MD: The Johns Hopkins University, Psychology Department.

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence*. New York: Columbia University, Teachers College, Bureau of Publications.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*, 433–451.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554.

Tiggle, R. B., & Rafacz, B. A. (1985). Evaluation of three local CAT-ASVAB network designs. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 23–28). San Diego, CA: Navy Personnel Research and Development Center.

Towle, N. J., & Merrill, P. F. (1975). Effects of anxiety type and item difficulty sequencing on mathematics test performance. *Journal of Educational Measurement, 12*, 241–249.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Urry, V. W. (1970). A monte carlo investigation of logistic test models. Unpublished doctoral dissertation, Purdue University, West Lafayette, IN.

Urry, V. W. (1977). Tailored testing: A successful application of item response theory. *Journal of Educational Measurement, 14*, 181–196.

Urry, V. W. (1981). *Tailored testing, its theory and practice. Part II: Ability and item parameter estimation, multiple ability application, and allied procedures* (NPRDC TR81). San Diego, CA: Navy Personnel Research and Development Center.

Vale, C. D. (1981, April). *Implementing the computerized adaptive test: What the computer can do for you*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10*, 333–344.

Vernon, P. E. (1979). *Intelligence: Heredity and environment*. San Francisco: Freeman.

Vicino, F. L., & Hardwicke, S. B. (1984, March). *An evaluation of the utility of large scale computerized adaptive testing*. Paper presented at the American Educational Research Association convention, New Orleans, LA.

Vicino, F. L., & Moreno, K. E. (1988, March). *Test-takers' attitudes toward and acceptance of a computerized test*. Paper presented at the American Educational Research Association convention, New Orleans, LA.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*, 213–217.

Wainer, H. (1983). On item response theory and computerized adaptive tests: The coming technical revolution in testing. *Journal of College Admissions, 28*, 9–16.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement, 26*, 191–208.

Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (1997). *Rescuing computerized testing by breaking Zipf's Law*. Ithaca, NY: Cornell University, Department of Statistics.

Wainer, H. (2000). Rescuing computerized testing by breaking Zipf's Law. *Journal of Educational and Behavioral Statistics, 25*.

Wainer, H., Bradlow, E., & Du, Z. (2000). Testlet response Theory: An analog for the 3-PL useful in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer-Nijhoff.

Wainer, H., & Braun, H. (2000). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–202.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*, 1–14.

Wainer, H., & Messick, S. (1983). *Principals of modern psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., Morgan, A., & Gustafsson, J-E. (1980). A review of estimation procedures for the Rasch model with an eye toward longish tests. *Journal of Educational Statistics, 5*, 35–64.

Wainer, H. & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*, 339–368.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wang, M. M. (1987, June). *Measurement bias in the application of a unidimensional model to multidimensional item response data*. Paper presented at an Office of Naval Research conference, Alexandria, VA.

Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. *Machine-Mediated Learning, 2*, 217–282.

Way, W. D. (1998) Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice. 17*, 17–27.

Way, W. D., & Steffen M. (1998, April). *Strategies for managing item pools to maximize item security*. Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.

Way, W. D., Steffen, M., & Anderson, G. S. (1998, September). *Developing, maintaining, and renewing the item inventory to support computer-based testing*. Paper presented at colloquium "Computer-Based Testing: Building the Foundation for Future Assessments," Philadelphia.

Wegner, T. G., & Ree, M. J. (1985). *Armed Services Vocational Aptitude Battery: Correcting the speeded subtest for the 1980 youth population* (AFHRL-TR-85-14). Brooks AFB, TX: Air Forces Human Resources Human Resources Laboratory Manpower and Personnel Division.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (Research Report 74–5). Minneapolis: University of Minnesota, Psychometric Methods Program.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473–492.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 81–129). Washington, DC: American Council on Education.

White, P. O. (1973). Individual differences in speed, accuracy, and persistence. In H. J. Eysenck (Ed.), *The measurement of intelligence* (pp. 246–260). Lancaster, England: Medical and Technical Publishing.

Wilbur, E. R. (1986). Design and development of the ACAP test item data base. *Proceedings of the 28th annual conference of the Military Testing Association* (pp. 601–605). Mystic, CT: U. S. Coast Guard.

Wilson, D., Wood, R., & Gibbons, R. D. (1984). *TESTFACT: Test scoring, item statistics, and item factor analysis*. Mooresville, IN: Scientific Software.

Wise, L. L., Curran, L. T., & McBride, J. R. (1997) CAT-ASVAB cost and benefit analyses. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Monographs, 3*, No. 16, 1–62.

Wohlwill, J. (1963). The measurement of scalability for non-cumulative items. *Educational and Psychological Measurement, 23*, 543–555.

Wolfe, J. H. (1985). Speeded tests—Can computers improve measurement? *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 49–54). San Diego, CA: Navy Personnel Research and Development Center.

Wolfe, J. H., Alderton, D. L., Cory, C. H., & Larson, G. E. (1987). Reliability and validity of new computerized ability tests. In H. Baker & G. J. Laabs (Eds.), *Proceedings of the Department of Defense/Educational Testing Service conference on Job Performance Measurement Technologies* (pp. 369–382). San Diego, CA: Navy Personnel Research and Development Center.

Wolfe, J. H., Alderton, D. L., Cory, C. H., Larson, G. E., Bloxom, B. M., & Wise, L. L. (1997). Expanding the content of the CAT-ASVAB: New tests and their validity. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 239–249). Washington, DC: American Psychological Association.

Wolfe, J. H., Moreno, K. E., & Segal, D. O. (1997). "Evaluating the predictive validity of the CAT-ASVAB." Chapter 18 (p. 175–180) in Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.) (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

Woodworth, R. S. (1910). Race differences in mental traits, *Science, 31*, 171–186.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Wright, D. (1987). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test* (RM-87-1), (pp. 1–27). Princeton, NJ: Educational Testing Service.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (1993). *BIMAIN 2: Item analysis and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific Software, Inc.

Zimowski, M. F. (1988, April). *The duplex design: An evaluation of the two-stage testing procedure.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Zimowski, M. F., & Bock, R. D. (1987). *Full-information item factor analysis of test forms from the ASVAB CAT pool* (MRC Report No. 87-1, Revised). Chicago: Methodology Research Center, National Opinion Research Center.

Zipf, G. K. (1949). *Human Behavior and the Principle of least effort.* Cambridge, MA: Addison-Wesley.

This page intentionally left blank