Foreword

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

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

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

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

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

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

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

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

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

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

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

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

Ronald K. Hambleton
University of Massachusetts at Amherst

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

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