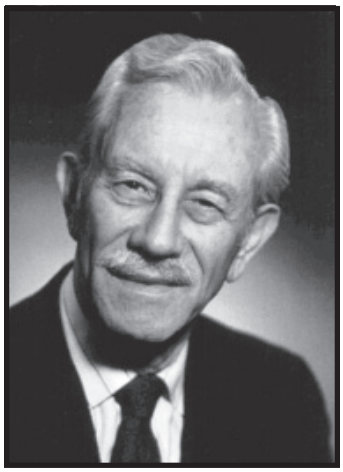
WILLIAM H. ANGOFF MEMORIAL LECTURE Series

THE SECOND CENTURY OF ABILITY TESTING: SOME PREDICTIONS AND SPECULATIONS

BY
SUSAN E. EMBRETSON

ETS Educational Testing Service
Research and Development
Policy Information Center

William H. Angoff
1919 - 1993

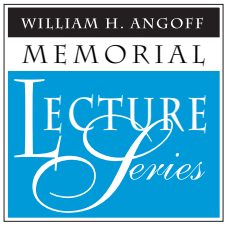


William H. Angoff was a distinguished research scientist at ETS for more than forty years. During that time, he made many major contributions to educational measurement and authored some of the classic publications on psychometrics, including the definitive text "Scales, Norms, and Equivalent Scores," which appeared in Robert L. Thorndike's Educational Measurement. *Dr. Angoff was noted not only for his commitment to the highest technical standards but also for his rare ability to make complex issues widely accessible.*

The Memorial Lecture Series established in his name in 1994 honors Dr. Angoff's legacy by encouraging and supporting the discussion of public interest issues related to educational measurement. The annual lectures are jointly sponsored by ETS and an endowment fund that was established in Dr. Angoff's memory.

The William H. Angoff Lecture Series reports are published by the Policy Information Center, which was established by the ETS Board of Trustees in 1987 and charged with serving as an influential and balanced voice in American education.

The Second Century of Ability Testing:
Some Predictions and Speculations



The seventh annual William H. Angoff Memorial Lecture was presented at Educational Testing Service, Princeton, New Jersey, on January 11, 2001.

Susan E. Embretson
University of Kansas
Department of Psychology

Educational Testing Service
Policy Information Center
Princeton, NJ 08541-0001

Preface

The ETS Policy Information Center is pleased to publish the seventh annual William H. Angoff Memorial Lecture, given at ETS on January 11, 2001, by Dr. Susan Embretson of the University of Kansas.

The William H. Angoff Memorial Lecture Series was established in 1994 to honor the life and work of Bill Angoff, who died in January 1993. For more than 50 years, Bill made major contributions to educational and psychological measurement and was deservedly recognized by the major societies in the field. In line with Bill's interests, this lecture series is devoted to relatively nontechnical discussions of important public interest issues related to educational measurement.

Dr. Embretson reviews the major developments in test theory, concepts, and methods that occurred during the 20th century—the first century of ability testing—and makes some predictions about developments that are likely to occur in this new century of testing. She predicts that the integration of cognitive theory together with advances in psychometrics will bring major, exciting changes in test development procedures, task design, and the range of abilities that are assessed. The research foundations and technology that will support these advances are being developed now and are likely to continue at an accelerating pace as we move into the second century of ability testing.

I believe that this lecture offers readers, regardless of technical background, a comprehensive introduction to the issues that have surrounded testing in the past and those issues that will be of concern in the future.

Drew Gitomer
Senior Vice President
ETS Research & Development
January 2003

Abstract

Dazzling changes in many areas, such as technology and communications, marked the 20th century—the first century of ability testing. Predictions about the second century of testing seem difficult in such a context. Yet, looking back to the turn of the last century, Kirkpatrick (1900) in his APA presidential address presented fundamental desiderata for ability testing (normative age standards, emphasis on culture-general tasks, simultaneous measurement of many persons, and adult ability measurement) that, in fact, guides major testing research even today. An overview of the last century shows that most fundamental principles in psychometrics and testing were available by 1930. With a few notable exceptions, the remainder of the last century of testing was devoted to applying or refining these principles. I predict that the same pattern will occur in this century of testing. Further developments in model-based measurement and cognitive psychology principles in testing, intermingled with technology, will guide ability testing throughout the next century. These changes, which I will elaborate in detail, include fundamental changes in test development procedures, the nature of the measuring tasks, aspects of ability that are measured, and types of interpretations given to ability.

Introduction

My purpose in this report is to glance into the second century of ability testing. Developments in test theory, concepts, and methods that occurred at the beginning of the 20th century—the first century of ability testing—remain influential in current testing practices. For example, the elaboration of true and error sources of test score variance is axiomatic to classical test theory (e.g., Spearman, 1904b), which remains the basis of most ability tests. But the 20th century marked progressively more dazzling changes in many areas, including areas that are seemingly related to testing, such as technology and communications. Looking backward, it is difficult to imagine that scholars at the turn of the 19th century could foresee such cultural mainstays as the automobile, jet plane, and Internet would replace the horse-drawn carriage and telegraph. Predictions about the second century of ability testing seem almost foolhardy in such a context.

Yet the future of ability testing may be less elusive if the past is examined intensively. That is, an examination of developments in ability testing that occurred during the 20th century may reveal trends that will continue in the future. In this report, I trace the foundations of ability testing from the turn of the 19th century to the end of the 20th century and present developments broadly, including construct development, testing issues, task design, test design, scoring models, psychometric methods, and evidence systems. Then I make predictions and speculations for the second century of ability testing, based on research in progress from the end of the 20th century.

The First Century of Ability Testing: A Brief Review

My purpose in reviewing the first century of ability testing is to find major trends and patterns that may aid in predicting the second century of testing. This review is not intended to be comprehensive. Interested readers should consult Thorndike and Lohman's review of the first century of testing (Thorndike & Lohman, 1990) for further information.

Precursors

Although not considered as part of the origins of ability testing, Francis Galton's anthropometric laboratory had a major influence on testing concepts. This laboratory, which was based in the South Kensington Museum in London, was an exhibit at the 1884 International Exposition. In the laboratory, individuals could be measured on a variety of low-level cognitive functions, including simple reaction time and performance of sensation and perception tasks, as well as physical traits such as hearing, muscular strength, keenness of vision, and the like. Although such measures today would not be part of ability testing, Galton seemingly believed in their value to measure intelligence. According to Galton (1883, p. 27), "The only information that reaches us concerning outward events appears to pass through the avenue of our senses; and the more perceptive the senses are of difference, the larger is the field upon which our judgment and intelligence can act."

More important than Galton's tests were his contributions to psychometric methods. Galton applied the normal curve to understand individual differences in functioning, and he also developed a statistic, the covariance, to represent relationships between measures. Pearson (1901) refined the covariance into a scale-free index of relatedness, the correlation, which is fundamental in test theory for establishing test reliability and validity. Perhaps most inspiring of Galton's contributions was the anthropometric laboratory itself; that is, the laboratory demonstrated that cognitive functioning could be measured objectively and evaluated systematically.

On the American front, James McKeen Cattell was an important figure in promoting the basic notion of ability testing. Cattell (1890) used the term "mental test" to characterize a series of tests that he was using to measure college students and others. Apparently inspired by Galton, Cattell also believed that intelligence could be measured from sensory and perceptual tasks. Like Galton, he collected objective and standardized measures of large samples.

Galton and Cattell's tests, however, are not usually regarded as the origins of contemporary intelligence testing. Interestingly, Galton's own statistical development, the covariance that was standardized by Pearson (1901), provided the necessary tool to falsify these tests as measures of intelligence.

The Turn of the Century

A conceptualization that would foretell the future of ability testing was presented just prior to the 20th century. Kirkpatrick (1900) in his APA presidential address presented fundamental desiderata for ability testing. According to him, ability tests developed for children should have the following properties: 1) normative

standards should exist for each age group, 2) the abilities should be tested so as to have equal opportunities to develop in children, 3) the tests are administered to whole classes or schools at one time, and 4) the tests or procedures are also applicable to adults. Although his speech preceded the often cited origins of modern testing (i.e., Binet & Simon, 1908; Yerkes, 1921), Kirkpatrick's conceptualization of the desiderata for ability testing characterizes ability testing programs even today. That is, normative age standards for scores, emphasis on tasks with experiential generality, simultaneous measurement of many children, and extension of the tests to adult ability measurement are major aspects of contemporary ability testing.

But Kirkpatrick's vision was not fulfilled by the tests of Galton or Cattell. The Pearson correlation coefficient provided just the means for Wissler (1901) to examine the viability of low-level tests of cognitive functioning to measure intelligence. Wissler found that these tests had very low correlations with each other and with a major criterion of learning, school achievement. In fact, grades in gym were better predictors of academic performance than the tests of cognitive functioning.

Major Developments

This section presents a selective review of the developments during the first century of testing. Several areas of development are covered: 1) individual intelligence tests, 2) group intelligence tests, 3) psychometric methods, and 4) concepts of intelligence and ability.

Individual intelligence tests. Alfred Binet is usually credited as the founder of modern intelligence testing, although the many significant precursors mentioned above also contributed to the foundations of testing. The single most important aspect of Binet's contribution is the use of higher-order cognitive tasks to measure intelligence. Specifically, tasks involving judgment, comprehension, and reasoning were the essence of the Binet-Simon scale (Binet & Simon, 1908; Binet, 1911) and Binet's global conceptualization of intelligence (Binet & Simon, 1905). The second most important aspect of Binet's contribution is the use of empirical criteria to select intelligence test items. Binet's two criteria were 1) task performance should increase with age and 2) task performance should be related to school achievement.

An interesting feature of Binet's system is that items and examinees are placed on a common scale. That is, both items and examinees are referenced to mental age. Tasks were scaled for mental age from empirical data on the performance of children at various ages. Examinees were scaled for mental age by their relative success in solving the age-calibrated tasks. Due to the age-calibrations of the tasks, examinees may be compared even if they do not receive the same items.

Binet (1911) believed that a diagnosis of retardation could be made by subtracting chronological age from mental age. Large negative values (e.g., 2 or more years) indicated retardation. Stern (1912/1914) refined the comparison by developing the IQ concept as a ratio (not a subtraction) of mental age to chronological age. The ratio IQ concept persisted in individual intelligence measurement until the mid 1960s. When the ratio IQ was replaced with normative scores, Binet's common scale measurement of items and examinees was de-emphasized. The common scale measurement of items and examinees, although conceptually interesting, was not easily integrated in the mainstream of psychometric methods, which consisted of classical test theory.

Terman (1916) was responsible for adapting the Binet-Simon scale for use in the United States. However, this was not a mere translation or the alteration of a few tasks. Terman not only added tasks; he also made some methodological changes. He not only standardized the directions and instructions; he also added a new criterion to item selection, namely internal consistency. Terman's work (Terman, 1916) resulted in the Stanford-Binet intelligence test, which remains a major individual intelligence test today as the Stanford-Binet IV. Although the test has changed over the decades (e.g., normative IQ scores replaced the ratio IQ scores), the current test remains remarkably similar to the early test.

Group intelligence tests. The Army Alpha and Army Beta tests are often cited as representing the beginnings of group intelligence testing. These tests were developed over the course of just a few months in the United States during World War I. The main goal was to classify or select the large number of recruits for military service. The Army tests were developed under the direction of Yerkes (summarized in Yerkes, 1921). Table 1 lists the item types that appeared on the Army tests. The tests were administered in paper and pencil format, with standardization in test administration procedures, instructions, and scoring. The verification, multiple choice, or simple completion format for the item types on the tests made scoring by a clerk feasible. Scores were interpreted by reference to empirical standards, which were represented (unfortunately) by letter grades, ranging from A to E. Data on the relationship of test scores to officer training and a variety of military criteria supported test validity (Yerkes, 1921).

Of course, the Army tests were not developed completely anew. As noted by R. M. Thorndike and D. F. Lohman (1990), standardized group testing had been underway in a variety of locations, including at Columbia by colleagues of E. L. Thorndike. Item types that were appropriate for group ability testing, administered in paper and pencil form, were developed prior to World War I. They included analogies, paragraph comprehension, sentence completion, information, block designs, and so forth. For example, a test developed by Scott (1913) not only included objective item types and norms, but also included crucial

Table 1 - Subtests for the Army Alpha and Army Beta Tests

Alpha subtests	*Beta subtests*
directions	mazes
arithmetical problems	cube counting
practical judgment	X-O series
antonyms	digit symbol
disarranged sentences	number checking
number series	pictorial completion
analogies	geometric construction
	information

validity data, namely high correlations of test scores with teachers' judgments on ability. The most directly relevant to the Army tests, Otis (1917) developed a test with verbal items that he contributed to the Army Alpha, while Pressey and Pressey (1918) developed nonverbal item types that provided a model for the Army Beta.

Aside from the item types and the test administration mode (i.e., group rather than individual testing), the psychometric model for these tests differed from the Binet-Simon scale (Binet & Simon, 1908) in several ways. Table 2 contrasts Binet's scoring with Yerkes and Anderson's point scale method (Yerkes & Anderson, 1915). The point scale differs from the Binet-Simon scale in item arrangement, scoring, and the basis of score interpretation. Thus, in the point scale, item types were administered in homogeneous blocks, credit was given for each passed, and, significantly, scores were interpretable by reference to group norms rather than to age. The normative basis for score interpretation provided a more reasonable interpretation of adult ability test scores.

Thus, the Army tests fulfilled more completely Kirkpatrick's vision (Kirkpatrick, 1900) than did Binet's tests. That is, applicability to adults and administration to large groups characterized the Army tests but not Binet's. Neither test had normative standards in the contemporary sense. The Army tests classified recruits in categories (i.e., letter grades) based on their relative scores, but these were related to mental age on the Binet (see Thorndike & Lohman, 1990). However, it is doubtful that the Army tests fulfilled Kirkpatrick's fourth desideratum, namely that the abilities tested have equal opportunity to develop, any better than did the Binet tests. For example, the Army Alpha subtests of Practical Judgment and Information have item content that is clearly dependent on specific cultural backgrounds. Other tests on the Army Alpha and most tests on the Army Beta probably do meet this fourth desideratum, however.

After World War I, educational testing for intelligence followed the basic model of the Army tests, using the point scale method. Homogeneous subtests with normative scoring became routine. However, the use of subtests resulted in interesting patterns of intercorrelations, not necessarily supporting a single general intelligence factor. Kelley's book, *Crossroads in the Mind of Man*, (Kelley, 1928) proposed specific abilities that corresponded to categories of the various item types, including spatial relationships and numerical and verbal facility, as well as memory and speed. This book

Table 2 - Binet-Simon Scale (1908) Versus Yerkes Point Scale (1915)

	Binet-Simon	***Yerkes***
Item arrangement	heterogeneous	homogeneous subscales
Scoring	pass/fail age criterion	credit for each item
Norms	age level	multiple populations

foreshadowed the major concern for the next several decades—the development of tests for more specific abilities. The development of multiple aptitude test batteries was especially spurred on by World War II, as recruits had to be selected for increasingly complex specialties.

Psychometric theory. Psychometric theory developed rapidly during the first part of the 20th century. Spearman (1904b, 1907, 1913) published a series of papers that developed fundamental aspects of classical test theory. Namely, Spearman introduced the concept of reliability and expanded its relationship to validity, true and error variance, and test length. The 1904 paper, for example, presents the now classic formula for correcting validity correlations for attenuation due to unreliability. This development required separating true from error variance in test scores. Spearman also proposed a formula for the impact of test length on reliability, which is known as the Spearman-Brown formula (Spearman, 1910).

Using internal consistency to select test items to improve reliability also appeared early in testing. Terman (1916) included internal consistency for selecting items for his revision of the Binet-Simon scale (which became the Stanford-Binet). Although the mainstay correlation for item analysis in classical test theory, the biserial correlation, had been developed quite early (Pearson, 1909), Terman apparently did not use it. Instead, he used groups categorized on total score as a criterion to determine if item-passing probabilities increased accordingly (Thorndike, 2002, personal communication). Formalization of methods to select items to improve internal consistency reliability, through item to total score correlations, was active in the 1930s (e.g., Zubin, 1934; Richardson & Stalnaker, 1933). However, the biserial correlation was probably applied to many group tests that followed the Army tests in the 1920s, given the assumption of biserial correlations in the papers of the 1930s.

Spearman, with a collaborator, Hart, pioneered the use of the pattern of correlations between a set of measures to determine the number of abilities (Hart & Spearman, 1912). The tetrad difference criterion could test if a single common factor (*g*, or general ability, presumably) could account for individual differences on the measures. If the tetrad differences were zero, then a single common factor was supported, but if not, then it was unclear how many common factors were needed. Thurstone (1931) generalized the tetrad difference rationale to multiple factors by successive evaluations of residuals after extracting additional factors in the centroid method of factor analysis. Although Thurstone's subsequent application of factor analysis to study multiple abilities (Thurstone, 1938) conflicted with Spearman's theory, in fact the method can be regarded as an extension of Spearman's method of using correlational patterns to understand intelligence.

Significant progress in scoring also occurred early in the 20th century. Kelley (1914) proposed that a more adequate scoring system for ability tests would result from normative standard scores. He proposed that *z*-scores be used to represent abilities. Otis (1917) refined these into more generalized standard scores, so the mean and standard deviation can be set to any arbitrary values. The standard score system, of course, remains current in ability testing today.

The scaling of item difficulties also received systematic attention quite early. Binet (1911) pioneered empirical methods to scale item difficulties in the mental age scale. Elsewhere, however, simple proportion passing (still a classical test theory mainstay) was applied to scale item difficulties.

The precursors of item response theory (IRT) are also found early in the 20th century. For example, matching examinees to a scaling of item difficulty was attempted on the CAVD, which tested item-completion, arithmetic, vocabulary, and direction-following abilities (Thorndike, Bregman, Cobb, & Woodyard, 1926). The rationale was that items at the examinee's level should have a probability of .50 of being solved. Thurstone (1925) had a more mathematical solution to this scaling, applying the phi-gamma hypothesis from scaling to ability measurement. Person and items were placed on a common scale by using the normal distribution to scale item-solving probabilities. According to Thurstone (1925, p. 436), "Each test question is located at a point on the scale so chosen that the percentage of the distribution to the right of that point is equal to the percentage of right answers to the test question for children." The population of examinees, of course, could be designated by *z*-scores. Thurstone (1925, p. 449) presented a graph of the resulting absolute scaling of items on mental ability. These transformations resulted in a common scaling of persons and items that is similar to that given by the normal ogive IRT model that was developed decades later.

IRT is regarded as having two distinct origins, Georg Rasch (1960) and Frederic Lord (Lord, 1953; Lord & Novick, 1968). From the 1970s onward, the measurement journals were flooded with articles generalizing early IRT models, developing new IRT models, and developing effective estimation procedures.

Interestingly, however, the impact of IRT on ability testing was quite limited at the end of the 20th century. Only a few large-scale tests had applied IRT by the late 1990s. The majority of psychological tests still were based on classical test theory, which was developed early in the 20th century.

Theories of intelligence. A viable theory of intelligence apparently preceded the actual development of effective measures. Spearman (1904a) proposed his two-factor theory of intelligence quite early. Although the Binet-Simon scale (Binet, 1911) did not follow from Spearman's theory, Spearman later regarded the heterogeneous collection of tasks in the Binet-Simon scale as highly consistent with his theory (see Thorndike & Lohman, 1990). That is, heterogeneous measuring tasks lead to a better reflection of *g*, general intelligence, because the specific factors cancel out. Spearman (1923, 1927) further elaborated his theories of intelligence and cognition prior to 1930.

In 1921, the proceedings of a symposium on the nature of intelligence were published in the *Journal of Educational Psychology* ("Intelligence and Its Measurement," 1921). The participants included major theorists and test developers of the time, such as Terman and Thorndike. The views were wide-ranging and included underlying factors such as judgment, learning, multiple abilities, *g*, and more.

After 1930, attention turned to multiple aptitudes, seemingly inspired by Kelley's theoretical elaboration of them (Kelley, 1928). Thurstone (1938), Guilford (1967), and many others developed theories and corresponding tests for the major abilities. Multiple aptitudes could be given a more rigorous test than what early theorists could have done, due to the development of principal factor and component analysis (Thurstone, 1931; Hotelling, 1933). Spearman's seemingly contradictory view of a single aptitude (Spearman, 1904a) was eventually integrated into a hierarchical framework with the multiple aptitudes. Theoretical organizations, such as those proposed by Horn (1968) or Carroll (1993), unify the theories through more sophisticated applications of factor analysis.

The nature and number of abilities were the major concerns of intelligence theorists until the late 1970s, when Sternberg (1977) published his componential theory of intelligence. The concern shifted to understanding the nature of intelligence by identifying the underlying cognitive processing involved in solving intelligence test items. Known as cognitive component research, this line continues today and has expanded to include many different item types that appear on tests. Carroll and Maxwell (1979) regarded cognitive component research as a fresh wind for intelligence research. This line of research is somewhat overshadowed by research that links intelligence to brain functions, which is made possible through imaging techniques. Although ability constructs are now often described by reference to cognitive processing, cognitive component research did not have direct impact on intelligence tests available in the late 1990s.

Sternberg and Detterman (1986) presented a contemporary group of intelligence theorists with the same questions that were given to the 1921 scholars on intelligence. Although some original views on the nature of intelligence had not persisted (e.g., instinctual basis) and some new ones had emerged (e.g., information processing metacomponents), Sternberg and Detterman found substantial similarity among the viewpoints across the decades.

Summary

Most of the fundamental principles in the nature of the measuring tasks, testing methods, psychometric theory, and theories of intelligence were available by 1930. First, the tasks required to successfully measure intelligence, judgment, and reasoning were found in Binet and Simon's individual intelligence tests (Binet, 1911) and in the Army Alpha and Army Beta, and they remain current in intelligence measurement today.

Second, general testing methods, including standardization of procedures and scoring, were clearly evident in the tests developed before 1920. Terman's standardization of the Binet-Simon scale and the Army Alpha and Army Beta testing procedures (Yerkes, 1921) provided the model for subsequent tests throughout the 20th century.

Third, most fundamental principles for psychometric methods were available by 1930. The conceptualization of reliability by Spearman (1904b), the development of appropriate statistics for item and test analysis (Pearson, 1901, 1909), a conceptual framework for factor analysis (Hart & Spearman, 1912; Thurstone, 1931), and an IRT-like common scaling of persons and items (Thurstone, 1925) were developed before 1930.

Fourth, the basic conceptualization of intelligence that guided subsequent testing for decades was in place by 1921. Later views clearly included some new aspects, but they did not differ radically from earlier views.

Oscar Buros (1977) described 1927 as the banner year when testing reached maturity. The foundation was laid for further developments. Indeed, these were exciting times!

Of course, further developments in all areas occurred in the middle and final decades of the 20th century. However, many developments were extensions or refinements of basic principles that were already available by 1930. As Thorndike and Lohman (1990) conclude in their review of the first century of ability testing, the pace slows down. Buros (1977) had a more extreme view. He regarded 1927 as the "banner year" when testing reached maturity, but believed that the 50 years thereafter resulted

in little new except for electronic test scoring and analysis. During these years, however, even into the 1990s, numerous publications were released that covered formalizing and collecting classical test theory (see Gulliksen, 1950), developing and elaborating IRT (e.g., van der Linden & Hambleton, 1996), developing factor analysis further, and doing more research on the number and nature of abilities (see Carroll, 1993, for a summary). The conceptual groundwork for these developments, however, may be traced to research prior to 1930.

Also, during this time period, the testing industry was very active, resulting in alternative scales for measuring individual intelligence (e.g., Wechsler, 1939) and a proliferation of aptitude tests. The testing industry became large and lucrative.

An informal survey of test catalogs at the turn of the 20th century (1999) shows that both individual intelligence tests and group tests follow the models that were established early in the century. For individual tests, similar item types are used and they are selected by the same general empirical criteria as Terman (1916). The major change after 1930 was the scoring system when normative IQs replaced age ratio IQs. For group tests, Yerkes' point scale method is employed, and many item types are similar to those that appeared on the Army Alpha or Army Beta. Classical test theory and normative scoring remain the predominant psychometric method.

The Second Century of Ability Testing

During the 1980s and particularly the 1990s, new principles for measurement were being formulated. These developments apparently prompted Bennett (1998) to hypothesize that testing would reinvent itself. He foresaw three generations of this reinvention that would differ in test purpose, content, format, delivery location, and technology. The first generation would consist of computer-based tests. These tests would have similar purpose and content as current paper and pencil tests, but would also have relatively small changes afforded by computer technology. The second generation would also consist of electronic tests, but the increasing impact of technology, cognitive science, and model-based measurement would change the content, development, and scoring of tests. The third generation, where testing reinvents itself, would consist of more radical changes. Bennett envisions testing as merging with instruction.

Like Bennett, I envision that changes are likely to develop from model-based measurement, cognitive analysis of items and tasks, and Internet delivery of tests. I predict that the following areas will change in the next 25 years: 1) test development procedures, 2) the nature of measuring tasks, and 3) the aspects of abilities that are measured.

Test Development Procedures

In the next 25 years, I anticipate these changes in test development procedures: 1) continuous test revision, 2) automated validity studies, and 3) item development by artificial intelligence.

Continuous test revision. In the first century of testing, test revision was a costly and time-consuming project. Discrete test forms were developed, and the forms needed an empirical tryout to establish test reliability, test validity (often correlation with the old test forms), and appropriate norms. Hundreds to thousands of examinees could be required for a revision, and sometimes compensation was required for access to the appropriate population. Revisions typically took 18 months to 2 years to complete; however, often they required effort over a period as long as 5 years. With this kind of time and expense, it is obvious why tests changed so little over the many decades of the last century.

However, I predict that the second century of testing will have continuous test revision. The revision will be implicit in the testing system itself. New items will be continuously calibrated relative to the old items and then automatically added to the test bank after minimum standards are met. That is, automated checks on item properties and fit can assure that items have sufficient quality to be permanently entered.

Such a system cannot be too far away. The Armed Services Vocational Aptitude Battery (ASVAB) testing system already has continuous item revision. New items are seeded into operational tests for administration in an adaptive testing system based on IRT. Although the calibration and evaluation of items are not yet automated, they could be, if the programs were linked and if target item parameters were specified.

A second aspect of continuous test revision is updating norms. Rather than restandardizing the test with a new version, if data collection is centralized, incoming new protocols could be the basis of updated norms. Of course, statistical sampling principles and case weights should be applied so that the norms remain representative of the same population.

The first requirement for continuous test revision is a centralized and computerized system of test delivery. Because large amounts of data need to be available quickly, a centralized and computerized delivery system seems essential. A second requirement is an invariant method of item calibration and ability estimation. That is, item parameter estimates must not be biased by the particular sample on which they are based. Since new items are seeded into the system at differing times, shifts in the examinee population could occur. IRT-based calibrations, fortunately, have the property of invariance that is required. Ability estimates also must be invariant over the particular items that are used and independent of norms. The possibly shifting population over continuous testing and with possibly differing items requires a method to place abilities on a common scale. Again, IRT-based ability estimates have this required property.

Automated validity studies. Like test revision, validity studies in the first century were costly and time-consuming. Criterion data or reference measures had to be collected, in addition to the test scores. Analysis was also time-consuming, requiring the merging of files and the application of appropriate statistics.

Two types of automated validity studies can be envisioned, which depend roughly on the distinction between construct representation studies and nomothetic span studies (Embretson, 1983). In construct representation studies, item properties, such as item difficulty and response time, are mathematically modeled from item stimulus features that represent cognitive processes. Such models not only elaborate the nature of the construct that is measured by the items, but they also have yielded adequate prediction of the psychometric properties for many item types (see Embretson, 1999). If test administration is centralized and computerized, feasible with Internet delivery of tests, item responses and response times would be continuously collected. A centralized system also could contain item stimulus features from cognitive models that predicted item psychometric properties. The model could be checked periodically to determine if the cognitive model held for the new data. Similarly, routine checks on item fit to a psychometric model (e.g., the IRT model) could be made. Consistency would indicate that the construct representation aspect of validity had been maintained and thus would provide further validity data. Inconsistency, on the other hand, could be used to troubleshoot sources of invalidity. For example, if an item fell below a certain criterion, it could be flagged for further checks or removal from the item bank. Bennett and Bejar (1998) envision that automated scoring for open-ended responses also could profit from similar ongoing evaluations.

The other aspect of construct validity, nomothetic span, concerns the relationships of test scores with external measures. The centralized test delivery system could also be organized to include other sources of information on examinees, such as criterion scores or school learning, demographic information, and other test scores. Analyses could include differential item functioning and the external correlates of test scores. Again, centralized programming could be developed to routinely assess nomothetic span with incoming data and compare it to previous results. As for the construct representation data, consistency provides new support for validity while inconsistency can be used to troubleshoot the test.

Like continuous test revision, automated test validity studies do not seem too far away.

Item development by artificial intelligence. If test items can be automatically calibrated and then adaptively selected for use, it is feasible at least to imagine a system in which new items could be written by a computer program. Although this seems rather futuristic, the precursors of such systems, in fact, are already in progress. Bejar (1990, 1996) describes an item generative testing system in which items are variations of "item models." The item model is an existing item that has satisfactory psychometric properties. Research to pilot item generation for mathematical items for the Graduate Record Examinations® (GRE®) is in progress. Embretson (1999) presented nonverbal reasoning items that were generated according to a cognitive model to target psychometric properties. These developments will be described later in "The Nature of the Measuring Tasks."

Item development by artificial intelligence has practical importance for adaptive testing. Adaptive testing requires large item banks with many items at all levels so that equally precise measurements can be obtained. Item generation also can have theoretical importance for construct validity, about which I will elaborate later.

It is only a slight leap to envision item generators as the source of new items for seeding into a continuously revised test. Such items could be evaluated automatically for fit and target psychometric properties prior to permanent entry into the item bank.

However, I envision an item generative testing system that goes beyond an item source for continuous test revision. If items can be created for seeding, they also could be created instantaneously for the examinee during testing. That is, new items are generated to target psychometric properties during the operational test. This vision, similar to Bejar's (Bejar, 1996), seems on the surface to conflict with basic measurement principles from the first century of testing. Calibrated items are essential to scoring. Hence, item development requires multiple stages and tryouts.

The resolution to this seeming conflict depends on what is calibrated. Rather than calibrate items, design principles can be calibrated. In turn, these calibrated design principles predict the psychometric properties of items. The requirements for this level of item generation are either a cognitive design system behind the items (see Embretson, 1998, 2001) or an item model (Bejar, 1996). In the former case, actual items are generated from deep structures that embed the cognitive design features for items. Several psychometric models that can include design features have been proposed, starting with Fischer's linear logistic test model (LLTM) (Fischer, 1973). I describe this later. In the latter case, new items are created as variations of old items with substituted stimulus features. In this case, data need to be given to support the psychometric calibrations for the item model as appropriate for the variants.

At the turn of the 21st century, several computer programs were developed that can generate items. ITEMGEN1 (2002) can produce six item types for nonverbal intelligence tests, including the matrix completion problems described below for measuring abstract intelligence. Other item generators are the Test Creation Assistant (Singley & Bennett, 2002) and the GRE math item generator (Bejar et al., in press). These generators are based on an item model, within which key features of the item are varied. For example, for math word problems, the specific number in the problem or the specific characters or

setting can be varied. These generators require that the substitutions do not change the difficulty from the original item that provided the model.

The development of item generators is time-consuming and somewhat expensive initially. Each item type requires its own cognitive design system, which is based on a separate research foundation. However, compared to the ongoing cost of human item writers, the practical feasibility probably will lead to a rapid expansion of item generators in the near future.

It is interesting to imagine a testing system that combines all three predicted aspects of test development procedures, continuous test revision, automatic and continuous validity studies, and item development by artificial intelligence. Such a system could be self-sustaining without human intelligence. I predict that such systems will be operational in the first quarter of the second century of testing.

The Nature of the Measuring Tasks

I predict several changes in the nature of the measuring tasks, including: 1) shorter and more reliable tests, 2) item generation by cognitive design principles, 3) greater use of essays, completions, and worked problems, 4) broad conceptualization of what constitutes a "test item," and 5) flexible mixtures of evidence for ability.

Shorter and more reliable tests. Classical test theory wisdom is that longer tests are more reliable. The Spearman-Brown prophecy formula predicts increased reliability as test length increases, assuming that items of equal qualities are added.

I predict that shorter and more reliable tests will soon become commonplace. Shorter and more reliable tests depend on adaptive testing, in which items are selected to provide optimal information about the examinee. IRT is used to equate scores over the differing sets of items. Tests today that are adaptive include the ASVAB, Test of English as a Foreign Language™ (TOEFL®), and the GRE.

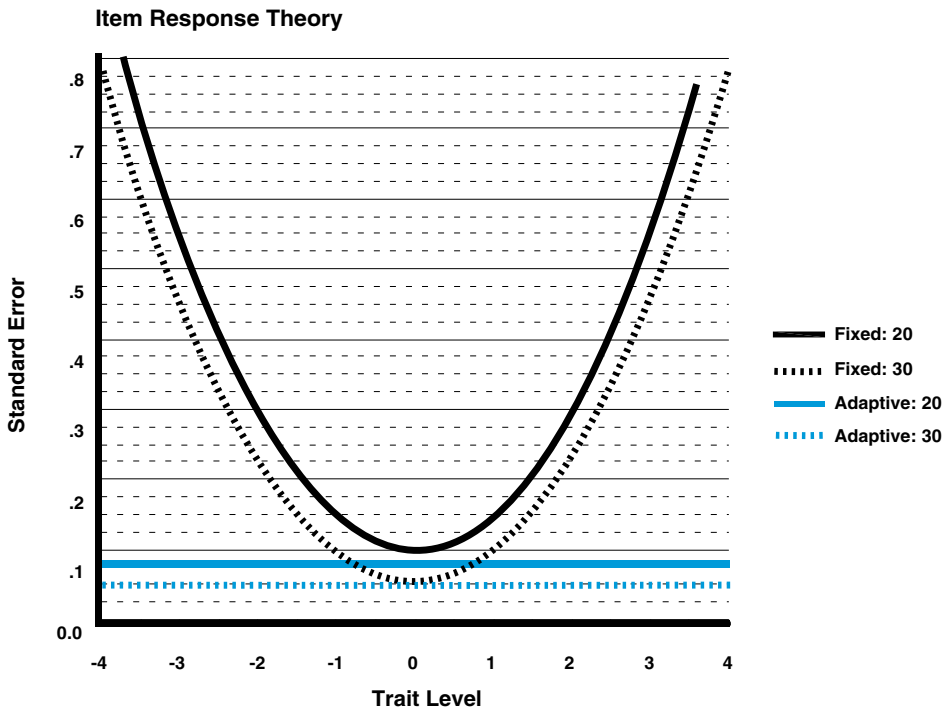
Figure 1 points out how both measurement error and test length can be reduced under adaptive testing. The standard error of measurement under IRT calibrations is shown for various ability levels for four tests from simulated data. The U-shaped curves display the standard errors under IRT calibrations for fixed length tests. In Figure 1, greater error is observed for estimating extreme abilities, as typical for fixed content tests, due to the fewer appropriate items for these examinees. Consistent with classical test theory, though, greater error is observed for all ability levels for the shorter (20-item) test than for the longer (30-item) test.

The other lines shown in Figure 1 represent standard errors for two adaptive tests from a large and wide-ranging item bank. In this case, measurement errors are approximately equal for all ability levels. And, consistent with classical test theory, the 30-item test has less measurement error than the 20-item test at all ability levels.

The important comparison is relative standard errors between the adaptive and the fixed content test. Notice that for most ability levels, the 20-item adaptive test yields less measurement error than the 30-item fixed content test. That is, we have a shorter and more reliable test. Obviously, the key to this effect is the selection of the most informative items in the adaptive test.

A possible incidental effect of shorter and more reliable tests is an impact on construct validity. Recently, I prepared two versions of an abstract reasoning test, a 34-item fixed content test and an 18-item adaptive test,

Figure 1 - Comparison of Measurement Error Between Four Tests With Varying Lengths and Testing Procedures



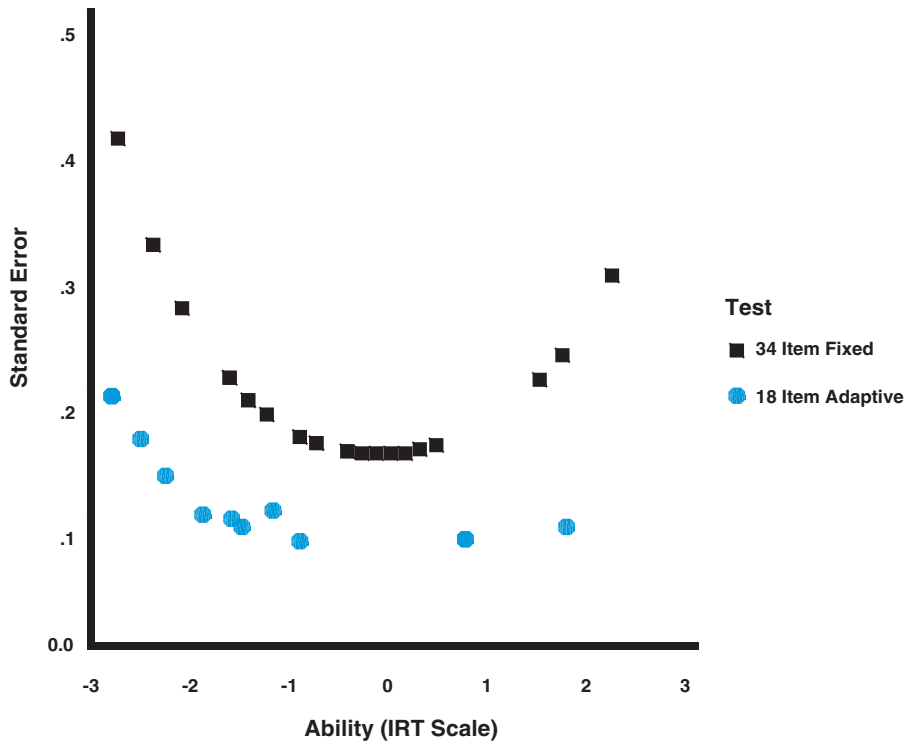
for a study on aging. One problem with ability measurement in an older population is that reduced motivation and self-efficacy may lower performance levels. Although 18 items are not many for measuring ability, the shorter and more appropriate test was also deemed by the investigator to maintain higher motivational levels. The results of a pilot sample on the two tests are presented on Figure 2. The 18-item adaptive test performs surprisingly well, with less measurement error at all ability levels than the fixed test, which had many items that were beyond the sample.

The intriguing issue about construct validity is this: Will performance levels of the elderly taking a shorter test increase relative to younger adults? If the lower performance of the elderly results from both their ability level and their lowered motivation, a shorter test may provide more valid measurement. Also, higher estimated abilities may result as well, which could change current wisdom about age-related declines in ability.

Item generation by cognitive design principles. In the first century of testing, it was almost axiomatic that test items must exist prior to test administration. Items were entered into either a test form or an item bank for use in adaptive testing. In the second century of testing, I predict that tests will no longer consist of existing items. Items will be written during the course of testing. That is, optimally informative items for measuring the examinee will be written instantaneously as needed by computer programs that are based on a deep theoretical understanding of item-solving processes.

At the turn of the century, research on cognitive component analysis of ability was extended to item generation (Embretson, 1998, 1999; Embretson & Gorin, 2001). A cognitive design system is based on cognitive mathematical models that predict item psychometric

Figure 2 - Measurement Errors for Two Tests on an Aging Population



properties and response times from their stimulus features. The stimulus features are linked to processing, such that each postulated process is represented by one or more stimulus features that control difficulty. Once the stimulus features are established in a mathematical model, the stimulus features of items may be manipulated to increase difficulty in the various cognitive processes. As item difficulty increases, item solving requires increased levels of the underlying cognitive ability.

A cognitive design system provides an effective method to select and display the specific stimulus features to be embedded in items for several reasons. First, items can be written for targeted difficulty levels, since the source of item difficulty is explicated by the cognitive model. Second, with a sufficiently powerful cognitive model, test items can be used without a tryout. The empirical properties of items are predicted by the cognitive model. Third, construct validity is obtained at the item level. The specific cognitive sources of item difficulty are known for each item. Fourth, full item generation by computer is feasible. Unlike the item-modeling approach to item generation described above, item structures need not be based on existing items, thus allowing new combinations of features. Fifth, large numbers of items can be generated quickly. Adaptive testing requires very large item banks for optimal measurement. Unfortunately, human item writers are unable to keep up with the demand for many tests. Sixth, greater test security may be possible, since the items need not even exist. That is, the specific item content is not needed prior to administration of the item. Only the design factors need be known.

An example of cognitive modeling research that leads to item generation is a series of studies on matrix completion items, which are used to measure abstract reasoning or general intelligence (Embretson, 1998). For

Table 3 - Theory for Matrix Completion Problems

Abilities	working memory capacity	abstraction capacity
Processes	goal management	correspondence finding
Item features	number of rules	abstract correspondence (rule level)

these items, Carpenter, Just, and Shell's theory of matrix solving was generalized to provide a mathematical model to predict item difficulty and response time (Carpenter, Just, & Shell, 1990). See Table 3. The theory postulates that the number and level of the rules in the matrix lead to increased goal management difficulty, which in turn requires larger working memory capacity.

Rule level, in contrast, also influences abstraction level that is required. Figure 3 shows a matrix completion item that has three rules and no abstraction.

The cognitive model for matrix completion items contains five predictors, a number of rules, abstraction, and three perceptual properties. Multiple correlations close to .80 are typically obtained from this model (see Embretson, 1998; 1999). For the item in Figure 3, prediction of item difficulty, ∃', is given by the following equation, where q_{ij} is the value of stimulus feature *j* in item *i*:

$$\exists' = .659q_{i1} + .792q_{i2} + \ldots\ldots\ldots - 1.719$$

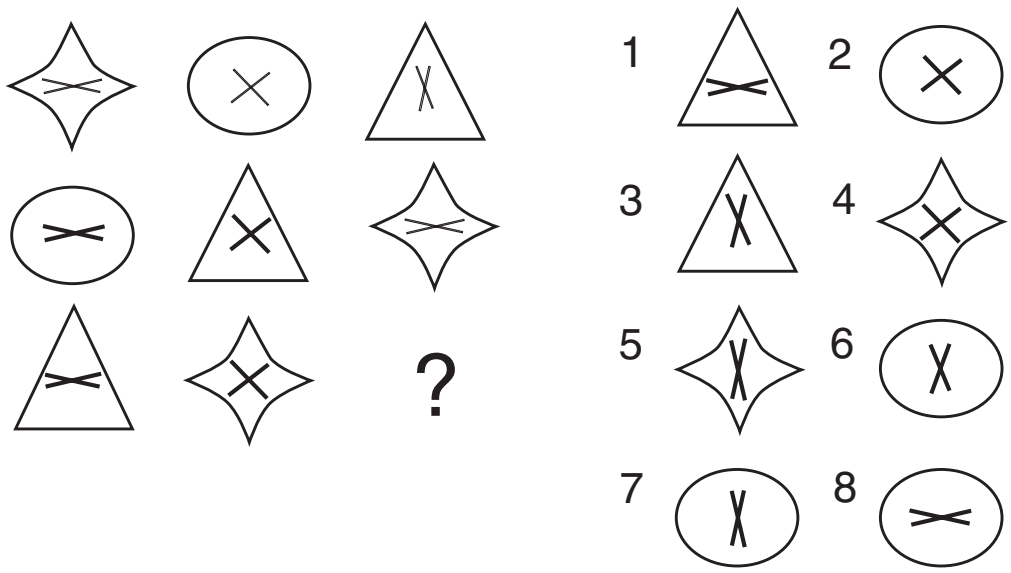
$$\exists' = .659(\#\ Rules) + .792(Rule\ Level) + \ldots\ldots\ldots - 1.719$$

$$\exists' = .255.$$

The model can be applied to any item that is produced. The model also can be used to produce items for targeted difficulty levels, with specific cognitive sources of difficulty. For example, an item with both working memory load and rule abstraction can be created by inserting stimuli into the matrix format, which leads to a high number of rules with high rule levels. For another example, developing an item with a large number of rules but with low rule levels can create an item in which only working memory load is important. The exact display of an item depends on item structure, in which the stimulus features are selected and displayed to fulfill the cognitive model (see Embretson, 1998, 1999).

Greater use of essays, completions, and worked problems. The first century of testing received great impetus from the development of item types that could be scored by stencils or (eventually) by electronic answer sheets. Large populations, such as recruits in World War I, could be readily tested. In the first century of testing, however, the objective item types that were available were limited to primarily multiple choice format. Other formats, such as essays, completions, and worked problems, required human raters, which led to greater expense,

Figure 3 - A Matrix Completion Item



unreliability, and delay of test scores. However, I predict that far greater use of essays, completions, and worked problems for measuring ability will occur relatively early in the second century of testing.

Some recent advances in automated scoring paves the way for using open-ended item formats. An early effort (Bejar, 1988) supported the potential of WordMap, an off-the-shelf program, to analyze grammatical errors in sentences, such as those that may occur in written tests. This effort apparently did not lead to a testing application, however. More recently, computer programs have been developed for scoring essays (Burstein et al., 1998) and graphical problem representations (Bennett, Morley, Quardt, & Rock, 2000).

The *e-rater™ system*, an automated essay-scoring program, mimics human rater's scores. The program models the raters' scores by scoring essays on a large number of linguistic variables, such as syntactic structure, vocabulary level, and word content. The raters' scores are regressed on the computer's linguistic scores to estimate optimal weights for prediction. Then, once the weights are estimated, *e-rater* is ready to score the remaining essays independently. The results on *e-rater* have been quite promising; for example, the correlation of *e-rater* scores with human raters has been found to be greater than the correlation of the human raters with each other.

Some caveats about the Burstein et al. (1998) approach, however, should be given. First, the human raters' scores that *e-rater* predicts may not have optimal validity. That is, the scores given by bleary-eyed raters after reading hundreds of essays may not reflect essay quality in the way that was intended (see Bennett & Bejar,

1998). Second, the *e-rater* approach is essentially atheoretical. The nature of the variables that provide optimal prediction of human raters is not a consideration in the weighting.

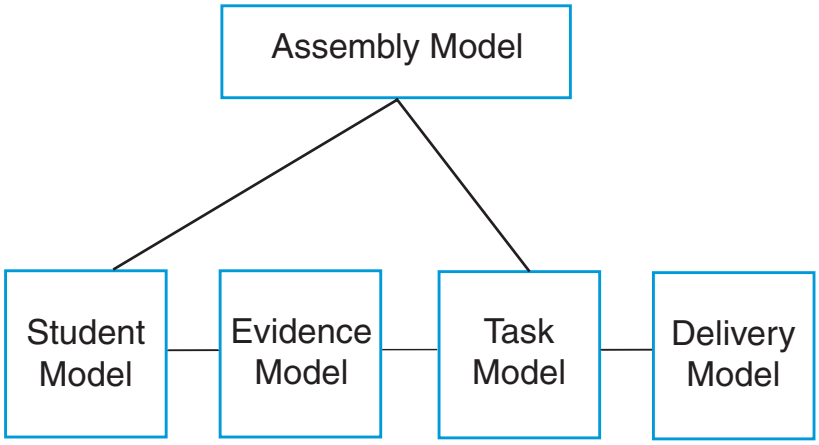
Bennett and Bejar (1998) describe a more theoretically driven approach to automated scoring. They envision an automated scoring approach that is primarily theory-driven, in that the scoring is intricately related to the construct definition, test design, and task design. In their approach, the features to be scored are selected and weighted according to a theoretical rationale. Unlike *e-rater*, the scoring reflects closely the intended validity of the test.

The developments in automated scoring of open-ended responses are exciting. With a bit more research and development, I predict that they will revolutionize the range of measuring tasks.

Broad conceptualization of what constitutes a "test item." In the second century of testing, I predict that rather unanticipated observations will have the role of measuring tasks. A broad conceptualization of measurement is given by Mislevy, Steinberg, and Almond (2001). In their evidence-centered approach, Mislevy, Steinberg, Breyer, Almond, and Johnson (1999) specify several models in the design of an assessment, as shown in Figure 4.

These models include a student model, an evidence model, and a task model. Most pertinent to the current discussion is the evidence model. In the evidence model, the salient features of a work product or other relevant behavior are extracted and summarized to determine

Figure 4 - The Evidence-centered Design for Assessment



the observable variables. Obviously, the term "work product" is quite a broad category in itself, and which features are to be extracted even further broadens the nature of measuring tasks. As summarized by Mislevy et al. (1999), the task of the evidence model is drawing inferences about what a student knows, can do, or has accomplished from limited observations of what a student says, does, or produces.

Although applications of the evidence-centered approach have been only illustrative so far (see Mislevy et al., 1999), the broad framework seems likely to be highly appealing for applications in the second century. However, the system is not practical unless statistical methods for combining flexible mixtures of evidence are available. This leads to the next prediction.

Flexible mixtures of evidence for ability. For the second century of testing, I predict that measurement of ability will involve flexible mixtures of evidence. Abilities may be estimated from a mixture of task success and qualitative aspects of performance. This prediction is not possible unless a method for model-based measurement is sufficiently broad to include diverse types of evidence. IRT, as currently postulated, is model-based measurement, but it does not seem sufficiently broad enough to capture the diverse sorts of evidence that may be presented. For example, the evidence may consist of a combination of essays, graphical drawings, solution paths in problems, efficient use of multimedia resources, and the course of instruction.

A sufficiently broad statistical framework is under development (see Almond, Steinberg, & Mislevy, in press; Mislevy et al., in press). Graphic modeling is a general framework for model-based measurement in that it subsumes IRT, latent class models, and factor analysis models. Priors can be incorporated into the system, such as prior knowledge about abilities, the item parameters, and task influences. The posterior, or the outcome, is the probable ability given the person's task responses and the priors.

Aspects of Ability That Are Measured

The changes I predict in the measurement of aspects of ability are: 1) the types of interpretations of ability scores, 2) the measurement of qualitative aspects of individual differences (e.g., processing strategies and knowledge structures), and 3) the measurement of modifiability of performance over changing test conditions.

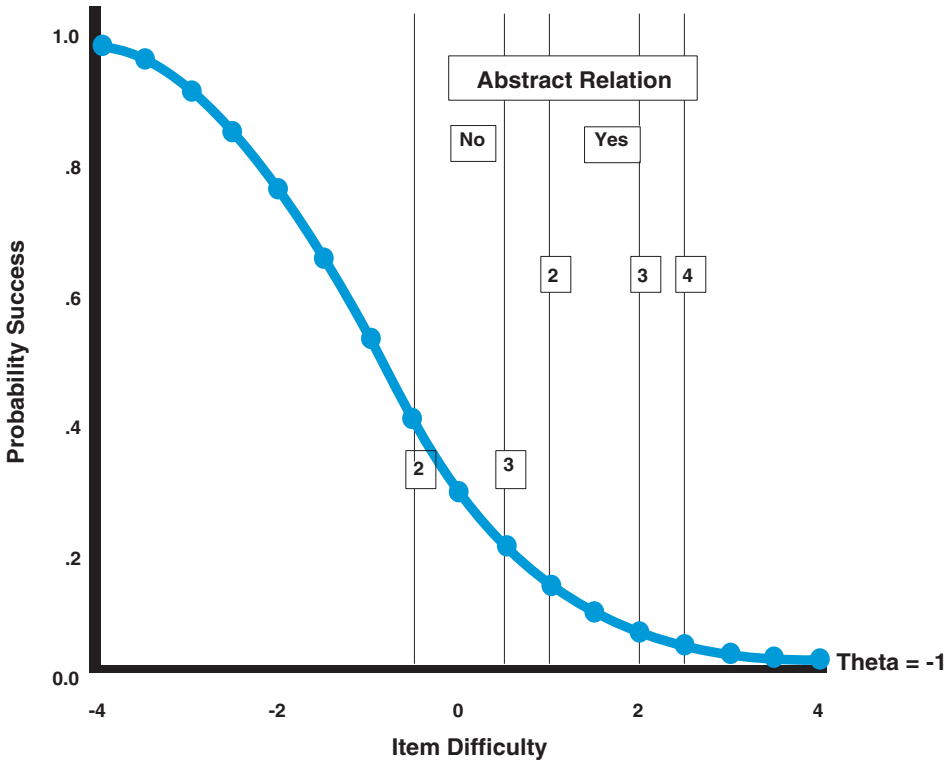
The types of interpretations of ability scores. Ability interpretations in the first century of testing were primarily normative. The examinee's score had meaning only in reference to the scores of other examinees. In contrast, (item) domain-referenced interpretations could be achievement test scores if subject matter experts stratified item content. Domain-referenced interpretations did not seem applicable to the relatively novel content of ability test items. However, the cognitive component research on ability from the last part of the 20th century gives rise to a new possibility. That is, abilities may be interpreted with reference to the processes, strategies, and knowledge structures that are involved in item solving.

Domain-referenced interpretations of ability require both a psychometric and a cognitive foundation. The psychometric foundation must be model-based measurement that includes indices for cognitive processing of items. In this case, common scale measurement would be obtained not only for items and persons, but also for the impact of cognitive processes on performance. Several IRT models, such as the LLTM (Fischer, 1973) and the 2PL-Constrained model (Embretson, 1999), have the required

property. Another approach (Sheehan, 1997) involves applying tree-based regression of cognitive properties on IRT calibrations. The cognitive foundation must be a plausible theory to link item psychometric properties to the stimulus features that underlie processing difficulty. Individual item types for measuring ability must be studied as cognitive tasks in their own right. In the first century of testing, the cognitive component studies of aptitude provide the beginnings of such plausible theories.

The enhanced domain-referenced interpretation of ability should note what processing the examinee can do easily and which processes are beyond him. An enhanced person characteristics curve, such as shown on Figure 5, illustrates domain-referenced interpretations. In a person characteristics curve, the probability for solving items of various difficulties is given for a person at a certain ability level. The example shown in Figure 5 is for matrix completion problems, as described above. Also shown in Figure 5 are locations on the item difficulty scale of Carpenter, Just, and Shell's major variables for cognitive processing, abstraction, and number of rules (Carpenter, Just, & Shell, 1990). These locations were obtained using a variant of tree-based regression to locate item categories. Locations are shown for abstract versus concrete relationships, as well as for the varying numbers of rules within the type of relationship. Given these locations, one can interpret the person's ability level by the probability that items with certain features can be solved. Figure 5 shows that the person has a moderate probability (about .40) of solving items with two rules when the relationships are not abstract. However, the person has a low probability (about .10) of solving items with two rules when the relationships are abstract.

Figure 5 - An Enhanced Person Characteristics Curve for Process-referenced Interpretations



Given the increasing demand by the test-taking public for more diagnostic testing, I predict that domain-referenced interpretations of ability will become prevalent. It should be noted, however, these interpretations are valid only for those persons whose patterns of performance fit the psychometric model. Although most persons will fit reasonably well, some will not. This leads to the next prediction.

Measurement of qualitative aspects of individual differences. In the first century of ability testing, a single aspect of ability was measured, namely, its level. However, it was often acknowledged that examinees also differ qualitatively so that the meaning of their ability scores differs. That is, examinees may differ in their patterns of processing competencies, in the strategies that they apply to solve items, in relevant background knowledge, in motivation, and in physical ways, such as handicaps and disabilities. These qualitative variants in item solving may render their ability scores incomparable. In the first century of testing, the main remedy was to determine whether or not these qualitative differences had impact on overall test validity.

In the second century of testing, I predict that these qualitative differences will be measured actively and used to guide score interpretations or to define moderator variables for the external correlates of test scores. Several psychometric developments published in the last part of the 20th century could provide the basis for measuring qualitative individual differences.

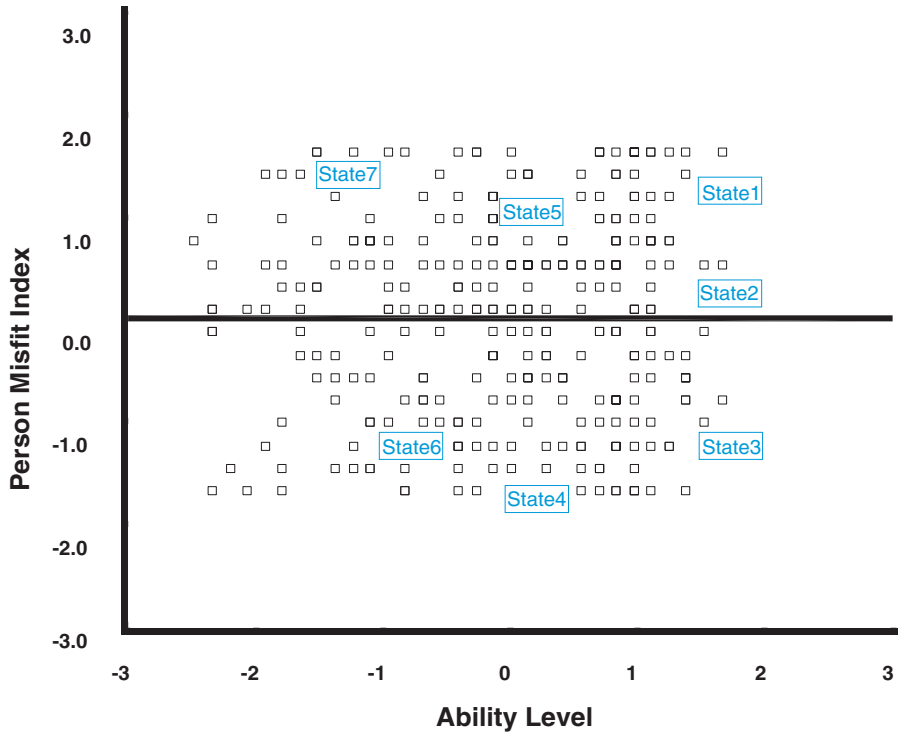
Person-fit indices (e.g., Drasgow, Levine, & McLaughlin, 1991) may be able to identify persons whose performance does not correspond to normative expectation. Person-fit indices may be estimated for tests that fit an IRT model reasonably well. Then a person-fit index may be estimated as the likelihood of their item responses, given the IRT model calibrations. A person's test protocol is unlikely if the normative order of item difficulty does not hold. Such a person would solve some very hard items but fail much easier items, for example. At the end of the last century, such fit indices were applied to check the validity of the person's ability scores (see Daniel, 1999).

However, qualitative differences among examinees may also be regarded as resulting from latent classes that differ in their approaches to item solving. Rule space analysis (Tatsuoka, 1985) and the mixed population Rasch model (Rost, 1990) assign examinees to latent classes based on their pattern of item responses.

In rule space analysis, the latent classes for item-solving are plotted in a two dimensional space defined by ability level and by a person-misfit index. Figure 6 shows a characteristic rule space, where the points represent examinees. Several latent states (classes) are also imposed in the rule space. These latent states are located in the plot from an ideal response pattern. For example, in an arithmetic problem, suppose an examinee does not know how to subtract when borrowing. To provide the ideal response pattern, each item is evaluated for requiring the rule: The item is scored pass if the rule is not required and scored fail if the rule is required. In turn, an ability and a misfit value are estimated for the ideal response pattern to locate the latent state. Examinees, then, may be classified into the latent state if their ability and misfit index is close to the latent state location. Figure 6 shows several latent states. An examinee's membership in a latent state could provide diagnostic information about the meaning of their ability.

The mixed Rasch model (MIRA) (Rost, 1990; von Davier, 1994) also can provide latent class membership for examinees. More than one latent class exists when more than one ordering of item difficulty is required to fit item response data for a population. Unlike rule space

Figure 6 - A Schematic for the Rule Space

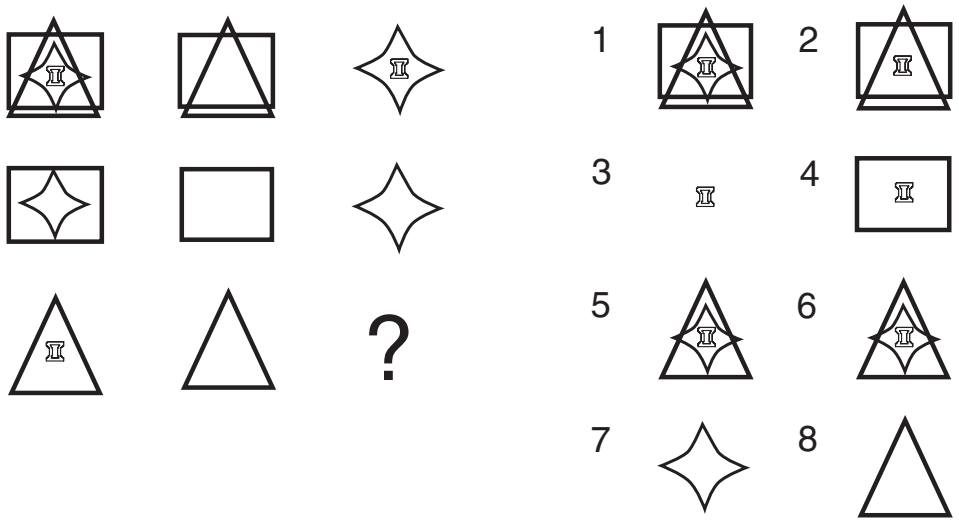


analysis, however, the latent classes are identified empirically in a data set. Each latent class is defined by a different ordering of item difficulty. In turn, item difficulty order is influenced by variables such as knowledge, processing strategy, processing component patterns, and other qualitative differences between examinees. An application of MIRA yields estimates of the number of latent classes, the item difficulties within each latent class, the proportion of the population within each class, and the individual class membership probabilities.

To illustrate, MIRA was applied to a large sample that had completed 34 matrix completion items. The Carpenter, Just, and Shell (1990) theory included a hierarchy of rules that provided the basis for determining whether or not abstraction was involved. Although it is assumed that all examinees apply the same type of rules to solve items, this may not be the case. Figure 7 shows a matrix completion item that can be solved by either the relatively easy holistic rule of figure addition/subtraction (i.e., values in the first and second column add up to produce the third column) or the harder analytic rule, the distribution-of-two rule. In the latter, two instances of each object occurs in a balanced fashion in each row and column. If an examinee does not know the holistic rule, the much harder analytic rule must be applied.

MIRA was applied to the data, and two classes were required to achieve fit. Figure 8 plots the item difficulty orders for the two classes. The regression line is plotted to show equal item difficulties in both classes. The items that may be solved by either rule are shown by circles. The figure shows that these items are much more difficult in Class 2 than in Class 1, thus supporting the existence of a class for which the easy rule is not known. Class 2 was

Figure 7- A Matrix Completion Problem That Can Be Solved by Two Difference Rules



large; approximately 39% of the population was estimated to belong to it. Persons in Class 2 also had lower levels of estimated ability than Class 1.

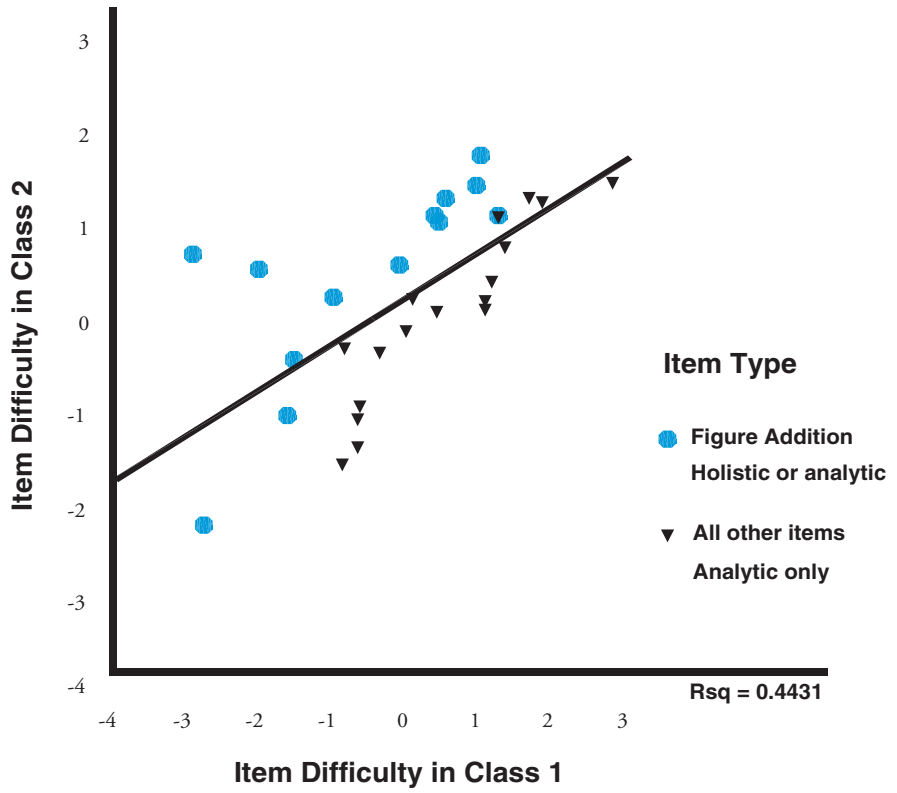
This leads to an interesting interpretative dilemma: Are abilities comparable between Class 1 and Class 2? In particular, if Class 2 had known about the easy holistic rule, would their scores be much higher? These questions are obviously central to the validity of the test as a measure of the construct for Class 2. If class membership is available with ability scores, we could use latent class as a moderator for predictions from test scores. Thus, the qualitative information could directly improve test validity.

For the second century of testing, I predict that qualitative differences in performance will be routinely assessed and interpreted. The models presented above, and perhaps new models, will be applied.

Measurement of performance modifiability over changing test conditions. I predict that performance modifiability, also known as dynamic testing, will increasingly provide the means to measure ability. In dynamic assessment, the responsiveness of the examinee's performance to cues, aids, instruction, or changing testing conditions is measured. Many different designs for dynamic testing are feasible; but a classic design includes a pretest, intervention, and a posttest.

Dynamic testing has been intriguing for several reasons. First, dynamic assessment is a seemingly more direct measure of learning potential, since learning itself may be included in the measurement design. Second, dynamic assessment may increase construct validity over static ability tests. That is, the instruction or cues provided may correct for preexisting differences in test

Figure 8 - A Scatterplot of Item Difficulties in Two Latent Classes



sophistication among examinees, thus making the posttest a more valid measure. Or the scope of prediction may be broadened by examining performance under varied conditions. Third, dynamic assessment may be useful in assessing concept mastery. That is, the task is presented under varying conditions, and presumably only those with the greatest mastery will succeed in the most challenging conditions.

As noted above, dynamic testing (e.g., Hamers, Sijtsma, & Ruijssenaars, 1993) was a topic of recurring interest in the last part of the 20th century. However, the psychometric basis of many tests was questionable, due to the use of unstandardized clinical procedures and the calculation of change scores with well-known problems (Harris, 1963).

Objective dynamic assessment has several requirements. First, item and cue selection must be not only adaptive, but also objective. That is, the next item or cue must depend on the person's responses, but the selection must be sufficiently objective as to eliminate human judgment. Second, item construction must be theory-based. Item stimuli and cues must be related to item solving processes. Further, item difficulty must be predicted from these stimuli. Third, comparable ability scores must be obtainable from different cues and items. This requirement almost requires IRT scaling to be effective and also, perhaps, partial credit scoring (e.g., Masters, 1982) to incorporate the impact of cues.

An example of an objective dynamic test is Guthke, Beckmann, and Dobat's (1993) Figure Series Test, a nonverbal reasoning test. This test exhibits many ideal properties for a dynamic measure. The test items were constructed on the basis of a theory (from structural information theory). The cues were administered adaptively, depending on item success. That is, if an item

were failed, then cues about the relationships were given. Ability estimates were adjusted for the number of cues given using a partial credit model. Full credit was given if the item could be solved with no cues and successively less credit for each cue administered. The properties of this test include computerizable item selection, partial credit scoring, theory-based item construction, and hierarchically ordered items and cues by the theory.

Objective dynamic assessment depends on further developments in other areas, such as plausible cognitive models for item solving and appropriate psychometric models (see Embretson, 1991). However, the needed components for objectivity are increasingly becoming available. I predict that dynamic assessment becomes a mainstay in ability testing. In fact, dynamic assessment may be the main mode of measurement, if instruction and testing merge as envisioned by Bennett (1998).

Summary

Although ability testing was relatively stable for the last several decades, I predict that the pace will increase sharply early in the second century of testing. Major changes in test development procedures, measuring tasks, and the abilities that are measured will occur at an accelerating rate early in the second century of testing. The many developments in progress at the turn of the century will be fueled by technology to lead to major changes. Much like the early decades of the last century, the early decades of the second century of testing will be exciting.

Test development procedures will evolve rather quickly in the technologically sophisticated society of the second century of testing. Continuous test revision, automated validity studies, and item development by artificial intelligence were predicted as major developments early in the second century of testing. The research foundations and technology required for these developments are in progress now and will probably accelerate rapidly.

The nature of the measuring tasks also will change quickly, with increasing Web-based delivery of tests and the employment of sophisticated model-based measurement methods. Tests will become shorter and more reliable quite soon. But not too far away are more drastic changes, such as item generation by cognitive design principles, greater use of essays, completions and worked problems, broad conceptualization of what constitutes a "test item," and flexible mixtures of evidence for ability. Measuring tasks will become increasingly flexible and may even include everyday behaviors (i.e., work products) as part of the measuring instrument.

Last, I predict that the aspects of ability that are measured will shift. Ability interpretations will be referenced to what the person can do, qualitative differences between examinees on the basis of their performance will be routinely measured and interpreted (e.g., processing strategies and knowledge structures), and performance modifiability (e.g., dynamic testing) will become a mainstay.

I envision these changes to occur rapidly and at an accelerating rate reminiscent of the first few decades of the 20th century. Most of the foundations for ability testing for the first century of testing were in place by 1930. And, like the first century of testing, the first few decades of the second century are predicted to be exciting times! But will the pace then slow down, after, say, 2030, as it did in the first century of testing? History would predict this.

With so many anticipated changes, it is important to revisit the essence of objective measurement. The basic psychometric principles that were pioneered early in

the first century will still be applicable, if conceived in a more general way. That is, objective measurement requires replication of behavior over tasks and conditions, empirical evidence on the psychometric properties of the tasks, and ability scores that depend on empirical calibrations of the measuring tasks. In the first quarter of the 20th century, these principles were developed and applied for the first successful intelligence tests. Diverse tasks were developed for measuring ability, tasks were calibrated on alternative basis (the Binet versus the point scale), and ability was linked to mental age or to norms. Although the application of the basis principles of objective measurement will differ in the second century, to generalize to the more flexible measuring tasks and aspects of abilities, the same basic principles will remain fundamental to testing.

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (in press). A four-process architecture for assessment delivery, with connections to assessment design. *The Journal of Learning Technology and Assessment*.

Bejar, I. I. (1988). A sentence-based automated approach to the assessment of writing: A feasibility study. *Machine-Mediated Learning*, *2*, 321-332.

Bejar, I. I. (1990) A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement*, *14*, 237-245.

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (ETS RR-96-13). Princeton, NJ: Educational Testing Service.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (in press). *A feasibility study of on-the-fly item generation in adaptive testing* (GRE Board Research Report 98-12). Princeton, NJ: Educational Testing Service.

Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing* (ETS Policy Report). Princeton, NJ: Educational Testing Service.

Bennett, R. E. & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, 9-17.

Bennett, R. E., Morley, M., Quardt, D., & Rock, D. A. (2000). Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning. *Applied Measurement in Education*, *13*, 303-322.

Binet, A. (1911). New investigation upon the measure of the intellectual level among school children. *L'Annee psychologique*, *17*, 145-201.

Binet, A., & Simon, T. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Annee psychologique*, *11*, 245-336.

Binet, A., & Simon, T. (1908). The development of intelligence in the child. *L'Annee psychologique*, *14*, 1-94.

Burstein, J. C., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., et al. (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT Analytical Writing Assessment essays* (ETS RR-98-15). Princeton, NJ: Educational Testing Service.

Buros, O. K. (1977). Fifty years in testing: Some reminiscences, criticisms and suggestions. *Educational Researcher*, *6*(7), 9-15.

Carpenter, P.A., Just, M.A. & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review*, 97.

Carroll, J. B., & Maxwell, S. (1979). Individual differences in ability. *Annual Review of Psychology*, 603-640.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

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

Daniel, M. H. (1999). Behind the scenes: Using new measurement methods on DAS and KAIT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement*. Mahwah, NJ: Erlbaum Publishers.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171-191.

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

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495-516.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 300-396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*, 407-433.

Embretson, S. E. (2001). Generating abstract reasoning items with cognitive theory. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development*. Mahwah, NJ: Erlbaum Publishers.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*, 343-368.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Galton, F. (1883). Inquiry into human faculty and its development. London: Macmillan.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

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

Guthke, J., Beckmann, J. F. & Dobat, H. (1993). Dynamic testing—Problems, uses, trends and evidence of validity. *Educational and Child Psychology, 14*, 17-32.

Hamers, J. H. M., Sijtsma, K., & Ruijssenaars, A. J. J. M. (1993). *Learning potential assessment: Theoretical, methodological and practical issues*. Amsterdam: Swets & Zeitlinger.

Harris, C. W. (Ed.). (1963). *Problems in measuring change*. Madison: University of Wisconsin Press.

Hart, B., & Spearman, C. (1912). General ability, its existence and nature. *British Journal of Psychology, 5*, 51-84.

Horn, J. (1968). Organization of abilities and the development of intelligence. *Psychological Review, 75*, 242-259.

Hotelling, H. (1933). Analysis of complex statistical variables into principal components. *Journal of Educational Psychology, 24*, 417-441.

Intelligence and its measurement. (1921). *Journal of Educational Psychology, 12*.

ITEMGEN1. (2002). *Item generator for non-verbal intelligence test items*. Lawrence, KS: Psychological Data Corporation.

Kelley, T. L. (1914). Comparable measures. *Journal of Educational Psychology, 5*, 589-595.

Kelley, T. L. (1928). *Crossroads in the mind of man*. Stanford, CA: Stanford University Press.

Kirkpatrick, E. A. (1900). Individual tests of school children. *Psychological Review, 7*, 274-280.

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

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

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

Mislevy, R. J., Senturk, D., Almond, R. G., Dibello, L. V., Jenkins, F., Steinberg, L. S., et al. (in press). *Modeling conditional probabilities in complex educational assessments* (CSE Technical Report). Los Angeles: Center for Studies in Evaluation, UCLA.

Mislevy, R. J., Steinberg, L. S., & Almond, R.G. (2001). On the role of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development*. Mahwah, NJ: Erlbaum Publishers.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (in press). On the structure of educational assessments. *Measurement: Multidisciplinary Research and Perspectives*.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R.G., & Johnson, L. (1999). A cognitive task analysis with implications for designing simulation-based performance assessment. *Computers in Human Behavior, 15*, 335-374.

Otis, A. S. (1917). A criticism of the Yerkes-Bridges Point Scale, with alternative suggestions. *Journal of Educational Psychology, 8*, 129-150.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine (Series 6)*, 2, 559-572.

Pearson, K. (1909). On a new method of determining correlation between a measured character A and a character B, of which only the percentage of cases wherein B exceeds or falls short of a given intensity is recorded for each grade of A. *Biometrika, 7*, 96-105.

Pressey, S. L., & Pressey, L. W. (1918). A group point scale for measuring general intelligence with first

results from 1,100 school children. *Journal of Applied Psychology, 2*, 250-269.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Richardson, M. W., & Stalnaker, J. M. (1933). A note on the use of bi-serial *r* in test research. *Journal of General Psychology, 8*, 463-465.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 3*, 271-282.

Scott, C. A. (1913). General intelligence or "school brightness." *Journal of Educational Psychology, 4*, 509-524.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333-354.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development*. Mahwah, NJ: Erlbaum Publishers.

Spearman, C. (1904a). "General intelligence," objectively determined and measured. American *Journal of Psychology, 15*, 201-293.

Spearman, C. (1904b). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5*, 417-426.

Spearman, C. (1923). *The nature of intelligence and the principles of cognition*. London: Macmillan.

Spearman, C. (1927). *The abilities of man*. New York: Macmillan.

Stern, W. (1914). *The psychological method of testing*. Baltimore: Warwich & York. (Original work published 1912).

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review, 31*, 356-378.

Sternberg, R. J., & Detterman, D. K. (Eds.). (1986). *What is intelligence? Contemporary viewpoints on its nature and definition*. Norwood, NJ: Ablex.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55-73.

Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin.

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence*. New York: Teachers College Bureau of Publications.

Thorndike, R. M., & Lohman, D. F. (1990). *A century of ability testing*. Chicago: Riverside Publishers.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*, 433-451.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review, 3*, 175-197.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, 1*.

van der Linden, W., & Hambleton, R. (1996). *Handbook of modern item response theory*. New York: Springer-Verlag.

von Davier, M. (1994). *WINMIRA. A program system for analyses with the Rasch model with the latent class analysis and with the mixed Rasch model*. Kiel, Germany: Institute for Science Education.

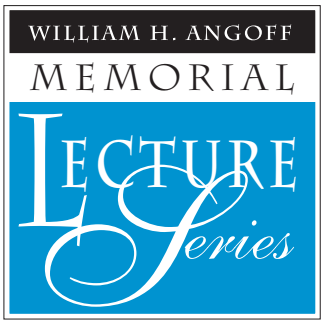
Wechsler, D. (1939). *The measurement of adult intelligence*. Baltimore: Williams & Wilkins.

Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Review Monograph Supplement, 3* (6).

Yerkes, R. M. (1921). *Memoirs of the National Academy of Sciences: Vol. 15. Psychological examining in the United States Army*. Washington, DC: National Academy of Sciences.

Yerkes, R. M., & Anderson, H. M. (1915). The importance of social status as indicated by the results of the point-scale method for measuring mental capacity. *Journal of Educational Psychology, 6*, 137-150.

Zubin, J. (1934). The method of internal consistency for selecting test items. *Journal of Educational Psychology, 25*, 345-356.



88503-006067 • S23M4 • Printed in U.S.A.
I.N. 996726