Journal of Computerized Adaptive Testing

Volume 11 Number 1
March 2024

The Influence of Computerized Adaptive Testing on Psychometric Theory and Practice

Mark D. Reckase
Michigan State University

DOI 10.7333/2403-1101001

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 11, Number 1, March 2024

DOI 10.7333/2403-1101001

ISSN 2165-6592

The Influence of Computerized Adaptive Testing on Psychometric Theory and Practice

Mark D. Reckase
Michigan State University

Editor's note: This paper is based on the author's keynote address at the 2010 meeting of the International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

This article presents a review of some of the issues that were prominent in the early days of large-scale assessment and presents an argument that the development of computerized adaptive testing (CAT) was influential in changing the way that these issues were addressed. It is argued that work on CAT increased the rate of adoption of item response theory (IRT) as the basis for scaling and reporting for large-scale assessments. It also shifted the emphasis of test analysis from the number-correct score to characteristics of test items and statistics related to those item characteristics. Further, the emphasis on interval scaling has declined because specifying the form of an item characteristic curve subsumes the assumption of an interval scale. These changes are the result of a paradigm shift in educational and psychological testing that resulted from the shift from classical test theory approaches to IRT-based approaches.

Keywords: *computerized adaptive testing, item response theory, scaling theory, paradigm shift, test design*

The major premise of this article is that part of the stimulus for the evolution of psychometric theory since the 1950s was the introduction of the concept of computerized adaptive testing (CAT) or its earlier non-CAT variations. The conceptual underpinnings of CAT that had the most influence on psychometric theory was the shift of emphasis from the test (or test score) as the focus of analysis to the test item (or item score). The change in focus allowed a change in the way that test results are conceived of as measurements. It also resolved the conflict among a number of ideas that were present in the early work on psychometric theory. Some of the conflicting ideas are summarized below to show how work on the development of CAT resolved some of those conflicts.

Psychometric theory first appeared as separate from statistical data analysis early in the 1900s. It is not the purpose of this article to document the beginning of psychometric theory, but a reasonable early example is Thorndike (1904) that included chapters on units of measurement, the reliability of measures, and sources of errors in measurements. My own training in psychometric theory did not require reading Thorndike. Rather, it started with classical test theory books such as Gulliksen (1950) and Nunnally (1967). I remember being impressed by a short book by Magnusson (1966) that seemed to be a concise summary of the thinking at that time about the theory of tests as producers of measurements. My training continued with a seminar based on Lord and Novick (1968) when the book was first published. It was interesting that the seminar did not reach Chapter 16 of that book titled “Latent Traits and Item Characteristic Functions.” Instead, the seminar focused on refinements of classical test theory. It was not until my academic advisor, Eric Gardner, handed me the book by Rasch (1960) when I was searching for a topic for my doctoral dissertation that I paid attention to the later chapters of Lord and Novick (1968).

Some Inconsistencies in Educational and Psychological Measurement Theory

I did not notice any inconsistencies in measurement theory until the oral defense for my master’s thesis (Reckase, 1971). The research for my master’s thesis was a comparison between two early methods for creating vertical scales for grade-level achievement tests—Thurstone’s (1925) absolute scaling method and Flanagan’s (1939) system of scaled scores. A key component of each of these procedures was assuming a shape for the score distributions for samples of students from adjacent grades taking the same form of a test. Thurstone’s method assumed normal distributions for scores at both grade levels, and Flanagan’s method relaxed that assumption but continued to use parametric forms for the distributions. When these methods were developed and used, they were considering score distributions based on number-correct scores but there was an assumption of underlying distributions on continuous scales. My research project was very limited in scope. I was investigating whether relaxing the normal distribution assumption made by Thurstone resulted in a difference in the vertical scale that was produced by the two methods.

After presenting my results, one of the faculty members on my committee asked me why it made sense to assume a normal distribution as the distribution of examinee proficiency underlying the observed score distribution. The question took me completely by surprise. Every course that I had taken in statistics and psychometrics, including those from that faculty member, made assumptions of a normal distribution for something. I had been taught that the assumption went back to the early work by Galton (1889) which showed that many of the measures that he took of people tended to have a normal (or Gaussian) distribution. Thorndike (1904) presented similar information and drew similar conclusions about the normal distribution. I tried that explanation in the oral defense, but the faculty member was not satisfied with that answer. He wanted me to give a more thorough justification for the assumption in my thesis. His demands required me to dig further into the theory of measurement.

Along with teaching about the use of the normal distribution in the development of parametric statistical procedures, my graduate program also covered measurement scale characteristics from the work of Stevens (1946) that indicated that an interval scale was needed to support most statistical procedures. Stevens also indicated that “Most psychological measurement aspires to create interval scales, and it sometimes succeeds.” (p. 679). Therefore, for the case of linking together grade-level achievement test scores, what was the argument for meeting the requirement for an interval scale?

One answer comes from an underappreciated connection between interval scale properties and the shape of a proficiency distribution. Unless an interval or ratio scale exists, the shape of a distribution does not exist. If a scale is assumed to have only ordinal properties, the scale can be stretched or compressed changing the shape of any distribution. That is essentially what the Thurstone and Flanagan scaling methods were doing. They assumed a distributional form was correct and did a transformation of the observed score distribution using a non-linear transformation of the summed-score scale to yield the assumed distribution. If the assumption of the distributional form was correct, the scale that resulted after the transformation to yield the desired shape was an interval scale. Rather than producing an interval scale and observing the shape of the distribution, they were forcing the data into a distributional form and observing the scale. The information from Galton, Thorndike, and Stevens seemed to give a good rationale for the use of the normal distribution as a basis for vertically scaling tests.

Another justification for the normal distribution assumption is the central limit theorem. The central limit theorem is described in many basic statistics texts as a justification for assuming the normal distribution for the sampling distribution of the mean. I first learned about this theorem in a graduate statistics course taught from Hays (1963). However, the exposition of the implications of that theorem in Hays' text was limited. Feller (1966) presented a much more thorough presentation of the central limit theorem (Chapter VIII, Section 4) which shows that the theorem applies to very general cases of sums of random variables. The implications are that sums of random variables tend to be normally distributed as the number of terms in the sum increases. That implies that the number-correct score, which is composed of the sum of a number of binomially distributed random variables, tends to form a normal distribution as the number of test items increases. The number of items required on a test for the central limit theorem to apply depends on the characteristics of the test items. The important implication for the purpose of this article is that approximating a normal distribution on the number-correct score scale implies that the scale meets the requirements for an interval scale.

Yet there were other publications in the psychology literature that dealt with methods for obtaining an interval scale. The most prominent of the publications, and the most technical, were the three-volume set of books written by various combinations of D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky. The first volume of that collection, Krantz, Luce, Suppes, and Tversky (1971), describes a particular measurement model called "extensive measurement". This measurement model is based on the concept of concatenation. The premise was that amounts on a scale can be obtained by adding entities together such as rods of equal lengths or objects of equal mass. The concept of concatenation is the basis for the measurement scales for length and mass (or weight). If test items can be assumed to indicate equal units of the construct being assessed, then the test could be argued to provide extensive measurement. The number-correct score would then have interval scale properties if all items were assumed to require equal amounts of the construct to attain a correct response.

Searching for the response to the question from the faculty member led me to another approach of defining an interval scale and the shape of a distribution. An obscure paper by Rozeboom (1966) developed the concept of scale from ordered equivalence classes. An equivalence class is a group of entities that are the same on the construct of interest. Ordered classes have all the entities in an equivalence class having more or less of the construct of interest than another equivalence class. Rozeboom indicated that entities in an equivalence class were not exactly equivalent on the construct of interest. Each class could be subdivided into smaller equivalence classes that contained more similar entities. At the limit, for a large population, as the subdividing of equivalence classes converged to single entries, the ordered classes approach a continuous scale.

I included this information in my thesis and satisfied the faculty member, and later published an elaboration of this discussion in a book chapter (Reckase, 1984). The chapter was titled "Scaling Techniques." Some reviews of the book indicated that this chapter was difficult to understand. Yet, it was included in the second edition of the book.

While this review of the literature on the properties of test scores was successful in the sense that I obtained the master's degree and published a book chapter, there was a nagging problem behind these different views on how to interpret the scores from tests. There are several arguments to support a normal distribution for modeling test scores. But there are also arguments for using summed scores or for assuming something other than a normal distribution to set the scale. It seemed that different scholars could scale the same test scores to create interval scales using different methods and obtain scales that were non-linear transformations of each other. How could there be an interval scale that is a non-linear transformation of another interval scale? Which of the scaling techniques gave the correct, or even the most preferred result? The conflicting results between scaling methods became even more evident when I started studying the Rasch (1960) model for use as a dissertation topic. The version of the Rasch model in his 1960 book was given as

$$P\left(u_{ij}=1\middle|A_jE_i\right)=\frac{A_jE_i}{1+A_jE_i}, \tag{1}$$

where u_{ij} is the score on the *i*th test item for person *j*,
A_j is the ability parameter for the *j*th person, and
E_i is the easiness parameter for the *i*th item.

For this version of the model, both *A* and *E* are on scales that range from 0 to positive infinity and they were described as ratio scales with true zero points. That is, an *A* parameter of 0 for a person indicated that they had 0 ability on the trait of interest yielding a 0 probability of answering any test item correctly. Similarly, an *E* parameter of 0 indicated that the test item had 0 easiness, meaning that it was so difficult that all examinees had a 0 probability of answering it correctly. These were very enticing properties of this model. A negative consequence of this model was that the distribution of estimated ability parameters for samples of examinees was very positively skewed—nothing like a normal distribution.

The version of the model proposed by Rasch (1960) is not the model that is now typically referred to in the measurement literature as the Rasch model. The most frequently seen model is

$$P\left(u_{ij}=1\middle|\theta_j,b_i\right)=\frac{\exp\left(\theta_j-b_i\right)}{1+\exp\left(\theta_j-b_i\right)} \tag{2}$$

where $\theta_j = \ln A_j$ is the ability parameter for person *j*, and
$b_i = -\ln E_i$ is the difficulty parameter for item *i*.

This model was proposed by Panchapakesan (1969) in her doctoral dissertation from the University of Chicago. She labeled it as the simple logistic model and stated that it was a special case of the logistic model proposed by Birnbaum (1968). She initially included another parameter in the model to indicate item discrimination but simplified the model to the current version by setting that parameter to 1.0. She particularly notes that the original Rasch model had a ”natural origin and definable unit for both of them [parameters].” When transformed into the model given in Equation 2, she noted that the origin of the parameters “is arbitrary.” However, she used the simple logistic model in her research because it supported convenient procedures for estimating the model parameters.

Even though the logistic form of the model has parameters that are a non-linear transformation of the parameters for the original Rasch model, some practitioners indicate that the logistic model provides parameters on an interval scale. For example, Fisher (1992) provided a demonstration of the scaling using the logistic Rasch model for scaling measures of distance. He concludes, “The overall linearity of the centimeter/logit plots demonstrates that data fitting the Rasch model produce interval scales equivalent to those in the physical sciences, just as Luce and Tukey (1964) called for.” (p. 51).

Therefore, if as Panchapakesan (1969) indicated, the original Rasch model has parameters on a ratio scale, Fisher (1992) indicated that the non-linear log transformation of those parameters produces an interval scale. How can that be? This summary provides many ways to justify assuming an interval scale for the reported scores from a test and many of them are inconsistent with each other. How can this state of confusion be resolved?

A Detour to Adaptive Testing

A defining characteristic of most adaptive tests is that each examinee is administered a test that is customized to their level of proficiency on the construct that is the target for the test (see Bandalos, 2018; p. 439, for a recent summary of CAT). There are two important features of the adaptive testing model. The first is that examinees take different sets of test items and sets of test items that are administered to each examinee cannot be considered as randomly parallel test forms. The items that are administered are purposely selected to vary in difficulty over examinees and for some test designs the *number* of test items administered might also vary over persons. The second is that adaptive testing focuses on test items, or sets of test items, rather than on the test as whole. The focus on test items is necessary for customizing the testing process to each examinee.

The focus on test items might seem like a trivial point in 2024, but in the 1970s most psychometric texts did not discuss the technical characteristics of test items, or if they did, item analysis was relegated to a later chapter in the book. The emphasis was on writing items to cover a domain of content. For example, the text by Stanley and Hopkins (1972) includes item analysis in Chapter 11 titled “Item Analysis for Classroom Tests.” The chapter covers simple procedures for computing percent correct and estimating the item-test correlation that can be used by classroom teachers. The focus is simple procedures for screening out poor items. This chapter comes well after chapters on “Test Validity” (Chapter 4) and “Test Reliability” (Chapter 5). Basic statistics are covered in Chapter 2. Nothing is mentioned about matching the difficulty of the test to the capabilities of the target population of examinees.

In the 1970s there was interest in customizing tests to each individual using procedures based on classical test theory. A good example is the set of procedures studied by Linn, Rock, and Cleary

(1969). They describe the challenges of estimating an examinee’s score when tests are adapted based on classical test theory. They describe many heuristics for producing scores, none of which were based on psychometric theory. Weiss and Betz (1973) provide a thorough summary of the research up to that year of the approaches to adapting tests to the examinee without using item response theory.

A Further Detour to Philosophy of Science

In a study of how changes occur in scientific thinking, Kuhn (1962) used the phrase “paradigm shift” to indicate when there has been a major change in the usual way of thinking about a topic. The classic example of a paradigm shift is changing the assumption that the Earth is the center of the Solar System to the assumption that the Sun is the center of the Solar System. Hellemans and Bunch (1988) suggest that Copernicus' publication of that change in perspective was the beginning of the scientific revolution (p. 91). That change in assumption resulted in major changes in the ways astronomers thought about the orbits of planets and the ways that they interacted with them. I am arguing here that CAT was instrumental in bringing about a paradigm shift in the way that psychometricians think about the way educational and psychological tests function.

Prior to the 1970s, the accepted way of thinking about test scores and test construction used true-score theory as the basis for understanding the accuracy of scores and how to develop scores that were equivalent across test forms. Texts such as Nunnally (1967) and Thorndike (1971) make passing mention of the normal ogive model and the parameters of the model, but conclude that the models are not very useful.

> “An important point to grasp in discussing monotone models with specified distributional forms [i.e., the normal ogive model] is that these models have not led to ways of scaling persons other than by the conventional approach, which is to sum scores on items.” (Nunnally, 1967; p. 72)
>
> “Until now, item characteristic curves have had little practical application in test construction. However, as high-speed computing facilities become more common, their attractive statistical properties may result in their wider use.” (Henrysson, 1971; p. 148).

However, the perspective on the use of IRT changed dramatically by the end of the 1970s. Yen and Fitzpatrick (2006) summarized the change in their chapter on IRT in the fourth edition of *Educational Measurement* (Brennan, 2006):

> “While IRT models were rarely used in education in the early 1970s, their use spread rapidly in the late 1970s and 1980s .... In the 1990s the use of the models spread, and they are now used in the majority of large-scale educational testing programs. In the past 15 years, use of IRT models has become commonplace for tests with selected-response (a. k. a. multiple-choice) items, constructed-response items, or combinations of selected-response and constructed-response items, for both paper-and-pencil testing and computerized adaptive testing.” (pp. 111-112).

The critical aspect of the use of IRT that resulted in the dramatic change in testing practices was the idea that estimates of examinee proficiency could be obtained on the same scale from any set of test items that had been calibrated on the same scale using an IRT model. When IRT was first developed as an alternative method for analyzing test data, almost all tests were based on the concept of a fixed test form with all examinees taking the same set of test items. When that form of test design is used, estimates of performance on the test based on true-score theory and IRT are highly related. In fact, for the Rasch IRT model, there is a functional relationship between the number-correct scores on a test and the proficiency estimate obtained from the model (see Fischer, 1995, for a thorough discussion of the connection). Because of that strong connection and the fact that it was much more difficult to perform IRT analyses than those based on true-score theory, there was slow adoption of IRT for the basis of test design and scoring. The paradigm shift in test theory was slow in occurring.

The events that led to the paradigm shift were the attempts by several organizations to implement CAT. Adaptive tests had existed for a very long time in the form of oral examinations of students, the development of the standardized version of the Stanford-Binet Intelligence Test (see Terman, 1916, for details) and other tests of that type and early attempts to adapt tests using true-score theory approaches. Many of these were individually administered tests using trained examiners or required complex administration procedures. A major problem was how to score the tests when different examinees responded to different sets of items. The traditional approach of summing the number of correct responses did not work because it did not account for differences in the characteristics of test items. The development of IRT provided an approach to scoring tests that was estimating the location on a construct using both the responses to the items and the characteristics of the test items. Bock and Mislevy (1982) indicate that Birnbaum (1958) was the "…first to show that maximum likelihood estimates of a person's ability can be obtained from an arbitrary set of items for which continuous response functions with respect to a common dimension can be specified." (p. 431). The IRT approach to scoring used powerful statistical methodology such as maximum likelihood methods or Bayesian procedures that were computationally intense, so their widespread use depended on the ready availability of computing power. That computing power became available with the development and marketing of personal computers.

Toward the end of the 20th century, a confluence of events brought about a paradigm shift in the theory underlying educational and psychological testing. Those events were the development of IRT models for tests, the development of readily available computers for the computation required to support IRT models, and the desire to customize tests for individuals to improve the accuracy and efficiency of the testing process. There was one other important requirement that led to the paradigm shift that was bringing together researchers and resources focused on developing a practical approach to computerized adaptive testing. That focused development was brought about through the funding and coordination from the staff of the Office of Naval Research (ONR). That organization funded applied research that was specifically focused on developing a CAT version of the Armed Services Vocational Aptitude Battery (ASVAB). An example of the collaboration that resulted from ONR funding and management is the extensive set of papers collected in Weiss (1977). The full history of the development of the CAT-ASVAB was documented in Sands, Waters, and McBride (1997).

Once the pieces were in place, there was a shift to using IRT as the basis for large-scale assessment programs. Software was available for calibrating test items using IRT models and for estimating examinees' locations on reporting score scales with different sets of calibrated items. The paradigm shift also highlighted the characteristics of test items as the basic components of tests because those characteristics were used when selecting items for administration by the CAT software. The test score was no longer the major focus of theory and analysis. Instead test item characteristics were the major focus and they were used in the test design and scoring process.

The Consequences of the Paradigm Shift

The consequences of the paradigm shift in educational and psychological measurement are apparent from noting the methodology used as the basis for most major testing programs in the world. For example, testing programs like NAEP, TIMSS, PISA, etc., are heavily dependent on item response theory and the capability to place estimates of performance on the same scale even though examinees do not receive the same set of test items (see Braun & von Davier, 2017, for a summary of the use of IRT for those programs). Further, most of the educational testing programs administered by states in the United States use IRT methods for scoring (estimating location on the scale), equating of test forms, and calibration of test items (Ryan & Brockmann, 2018). Likewise, many certification and licensure tests used for entry to the professions use IRT methodology (Way & Gialluca, 2017). A review of testing programs in the 1970s and 1980s would not identify testing programs that used IRT. The transition from methods based on true-score theory to methods based on IRT was slow, but the use of IRT methods is now dominant. It is not that the conceptual framework from true-score theory has been abandoned; it has been enhanced keeping relevant ideas such as true score and standard error of measurement but has used those ideas with scales defined using IRT models.

The shift to the use of IRT models has brought the field closer to the ideal of having estimates of proficiency on interval scales, although, interestingly, discussions of scale types have almost disappeared from the testing literature. For example, an article by Flake and Fried (2020) that is critical of the methods used to develop measurement instruments in psychology does not mention scale types nor does it reference any of the literature related to interval scaling. Perhaps the reason for the lack of discussion of scale types is that the selection of an IRT model requires the assumption of an interval scale. The shape of an item response function (IRF) for an IRT model does not exist unless the θ scale is assumed to be an interval scale. In effect, the assumption of the form of the IRF has replaced the assumption of an interval scale. The interval scale is a consequence of the assumption of a form for the IRFs, and scale properties are supported if there is evidence of fit to the proposed model.

The alternative to making the IRF shape assumption is to use something like Mokken scaling that uses ordering of items and proficiencies, but only specifies that the probabilities of correct responses are monotonically increasing with the increase in the order of proficiency of examinees (Mokken & Lewis, 1982). However, there have been only a few articles that show the use of Mokken scaling for tests of aptitude or achievement. Almost all operational tests of proficiency now use IRT methodology to define their reporting score scales.

The fact that using IRT for describing the functioning of test items and the locations of persons on the estimated reporting scale allows flexibility in the selection of test items while still reporting results on a common scale has freed up the test design process and facilitated the development of CAT. The major change is that IRT supports designing tests to maximize information or minimize standard error for the inferences that are desired from the test scores. A desired inference might be identifying the location of a person on the target construct for the test or making a decision about whether an examinee is above or below some critical point on that construct. Under the classical test theory paradigm, tests were designed to maximize measures of reliability or the correlation of test scores with a criterion measure. The standard error of measurement was usually assumed to be the same for all examinees. The IRT-based test design paradigm emphasizes building a test plan to provide the amount of information needed at various points along the scale defined to represent the construct and it does not require that all examinees respond to the same test items as has been demonstrated by CATs.

The paradigm shift is not yet complete. There are many IRT-based reporting score scales that are given labels which suggest that they are giving information about common constructs, such as "reading", but there is little evidence to suggest that the constructs are the same on those tests, or that they are intended to be the same. For example,

> "The NAEP reading assessment framework defines reading as a dynamic cognitive process that involves understanding written text, developing and interpreting meaning, and using meaning appropriately for text type and purpose.
>
> - The NAEP reading framework specifies the use of literary and informational texts to measure students' comprehension skills. The proportion of literary and informational texts varies by grade, with a greater proportion of literary texts at grade 4 and a greater proportion of informational texts at grade 12.
> - Students read grade-appropriate texts reflecting many content areas and respond to both selected-response and constructed-response questions about the texts they read. By design, the texts used in the assessment require interpretive and critical skills. The reading skills assessed are those that students use in all subject areas in school as well as in their out-of-school reading."
>
> NAEP Reading: Reading Results (nationsreportcard.gov) (National Center for Education Statistics, 2022)

is the way that the construct for the NAEP reading assessment is described. In Michigan, the Department of Education (2002) defines reading, supported by the International Reading Association, as

> "… the process of constructing meaning through the dynamic interaction among the reader's existing knowledge, the information suggested by the written language, and the context of the reading situation."

Although both definitions indicate that reading is a dynamic process, the details of that process are different. The NAEP definition emphasizes text types while the Michigan definition includes the interaction of texts with existing knowledge. The implication is that the constructs defined by the sets of test tasks selected for each kind of assessment are somewhat different.

Because of these subtle, and not so subtle, differences in construct definitions, tests that operationally define similar constructs define different reporting score scales rather than attempt to develop common constructs. The result is that educational and psychological measurement has not yet reached the quality of measurement enjoyed by the physical sciences where constructs of length, mass, time, etc. have common meaning even if they sometimes use different numerical scales. The different numerical scales are functional transformations of each other. No one argues that length measured in Michigan is different than length measured in Florida.

The ideal would be to have consensus definitions of constructs that are used for common educational and psychological concepts, and that all tests related to those constructs provide estimates of location on common scales. The methodological developments in item response theory and computerized adaptive testing can support those efforts. Large, calibrated item banks can provide tools for gaining information about individuals' locations on scales while still allowing customization of item selection. This is already being done in many testing programs that use content balancing methodologies within CAT testing procedures (Parshall, Spray, Kalohn, & Davey, 2002). Content balancing can have two functions. One is to estimate the commonly defined construct. The second is to customize the content.

It is unfortunate that there is still confusion about the intended meaning of the reported scores from tests. In the popular press, or in common usage, measures of reading proficiency are considered as indicators of the same construct when they come from different assessment programs such as NAEP, PISA, state assessments, or tests from commercial publishers. There is a desire to compare the results from these varied assessment programs. Yet, those who are responsible for producing those tests do not consider them equivalent because they require custom development of test items and tests for each application. A solution is needed for this complex problem. The language for describing test constructs is not precise, nor is the methodology for designing and constructing tests. There is a need to place the assessment of common constructs in a wider framework that can account for the slight differences in construct definitions and test content while still using a common cognitive space to describe the performance of examinees. It took from the 1960s until the 2000s to shift from a classical test theory framework to one based on item response theory. It might take an equal number of years to complete the paradigm shift.

References

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences.* The Guilford Press: New York.

Birnbaum, A. (1958). *On the estimation of mental ability* (Series Report No. 15, Project No. 7755-23). USAF School of Aviation Medicine: Randolph Air Force Base, TX.

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

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431-444. CrossRef

Braun, H. & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: psychometric and statistical considerations. *Large-Scale Assessments in Education, 5,* 1-16. CrossRef

Brennan, R. L. (2006). *Educational measurement (4th edition).* American Council on Education and Praeger: Westport, CT.

Feller, W. (1966). *An introduction to probability theory and its applications. Volume II.* John Wiley & Sons: New York.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer and I. W. Molenaar (Eds.) *Rasch models: Foundations, recent developments, and applications.* Springer-Verlag: New York. CrossRef

Fisher, W. P. Jr. (1992). Objectivity in measurement: A philosophical history of Rasch's separability theorem. In M. Wilson (ed.) *Objective measurement: Theory into practice.* Ablex Publishing, Norwood, NJ.

Flake, J. K. & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 446-465. *CrossRef*

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

Galton, F. (1889). *Natural inheritance.* Macmillan: London. *CrossRef*

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

Hays, W. L. (1963). *Statistics for psychologists.* Holt, Rinehart and Winston: New York.

Hellemans, A. & Bunch, B. (1988). *The timetables of science: A chronology of the most important people and events in the history of science.* Simon & Schuster: New York.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (ed.) *Educational measurement (2nd ed.).* American Council on Education: Washington, DC.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement: Volume I: Additive and polynomial representations.* New York: Academic Press.

Kuhn, Thomas (1962), *The structure of scientific revolutions*. Chicago, IL: The University of Chicago Press.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement, 29*(1), 129–146. *CrossRef*

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

Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. *Journal of Mathematical Psychology, 1*(1), 1-27. *CrossRef*

Magnusson, D. (1966). *Test theory.* Addison-Wesley: Reading, MA.

Michigan State Board of Education (2002). *Certification standards for the preparation of all secondary teachers, reading instruction.* Michigan Department of Education: Lansing, MI.

Mokken, R. J. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous responses. *Applied Psychological Measurement*, *6*, 417–430. *CrossRef*

National Center for Education Statistics (2022). *NAEP report card: Reading.*

Nunnally, J. C. (1967). *Psychometric theory.* McGraw-Hill: New York.

Panchapakesan, N. (1969). *The simple logistic model and mental measurement* [Unpublished doctoral dissertation]. University of Chicago.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer. *CrossRef*

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

Reckase, M. D. (1971). *A comparison of two methods for scaling test scores* [Unpublished master's thesis]. Syracuse University.

Reckase, M. D. (1984). Scaling techniques. In G. Goldstein and M. Hersen (Eds.) *Handbook of psychological assessment.* Pergamon: New York.

Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. S*ynthese, 16*(2), 170–233. *CrossRef*

Ryan, J. & Brockmann, F. (2018). *A practitioner's introduction to equating with primers on classical test theory and item response theory.* Council of Chief State School Officers: Washington, DC.

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

Stanley, J. C. & Hopkins, K. D. (1972). *Educational and psychological measurement and evaluation.* Prentice-Hall: Englewood Cliffs, NJ.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680. CrossRef

Terman, L. M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale.* Houghton Mifflin: Boston. CrossRef

Thorndike, E. L. (1904). *An introduction to the theory of mental measurements.* The Science Press: New York. CrossRef

Thorndike, R. L. (1971). *Educational measurement (2nd ed.).* American Council on Education: Washington, DC.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*(7), 433–451. CrossRef

Way, W. D. & Gialluca, K. A. (2017). Estimating, interpreting, and maintaining the meaning of test scores. In S. Davis-Becker and C. W. Buckendahl (eds.) *Testing in the professions: Credentialing policies and practice.* Routledge: New York. CrossRef

Weiss, D. J. & Betz, N. E. (1973). *Ability measurement: conventional or adaptive? Research Report 73-1*, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis. WebLink

Weiss, D. J. (1977). *Proceedings of the 1977 Computerized Adaptive Testing Conference.* Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis. WebLink

Yen, W. M. & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (ed.) *Educational measurement: Fourth edition.* American Council on Education and Praeger: Westport, CT.

Citation

Reckase, M. D. (2024). The influence of computerized adaptive testing
on psychometric theory and practice,
Journal of Computerized Adaptive Testing, 11(1), https://doi.org/10.7333/2403-1101001

Author Address

Mark D. Reckase
Email: reckase@msu.edu