EDUCATION
RESEARCH SERVICE

REVIEW OF THE LITERATURE

A primer on standardized testing: *History, measurement, classical test theory, item response theory, and equating*

Igor Himelfarb, PhD

Objective: This article presents health science educators and researchers with an overview of standardized testing in educational measurement. The history, theoretical frameworks of classical test theory, item response theory (IRT), and the most common IRT models used in modern testing are presented.
Methods: A narrative overview of the history, theoretical concepts, test theory, and IRT is provided to familiarize the reader with these concepts of modern testing. Examples of data analyses using different models are shown using 2 simulated data sets. One set consisted of a sample of 2000 item responses to 40 multiple-choice, dichotomously scored items. This set was used to fit 1-parameter logistic (PL) model, 2PL, and 3PL IRT models. Another data set was a sample of 1500 item responses to 10 polytomously scored items. The second data set was used to fit a graded response model.
Results: Model-based item parameter estimates for 1PL, 2PL, 3PL, and graded response are presented, evaluated, and explained.
Conclusion: This study provides health science educators and education researchers with an introduction to educational measurement. The history of standardized testing, the frameworks of classical test theory and IRT, and the logic of scaling and equating are presented. This introductory article will aid readers in understanding these concepts.

Key Indexing Terms: Chiropractic; Education; Educational Measurement

J Chiropr Educ 2019;33(2):151–163 DOI 10.7899/JCE-18-22

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

INTRODUCTION

In the 20th century, the concept of public protection dictated implementation of licensing laws to those professions having a direct relationship to public health and safety.[1] A plethora of discipline-specific prelicensure standardized assessment instruments (tests) exists to ensure compliance with the disciplinary standards. In the chiropractic profession, every year thousands of students take the prelicensure Part I, II, III, and IV examinations of the National Board of Chiropractic Examiners. As with any examination, some students feel that these standardized tests are unfair and have little relevance to clinical practice. Even faculty members often understand little about the boards. This article aims to provide an introduction to the world of standardized assessment not only for chiropractic educators but also for any health sciences educator or educational researcher.

OVERVIEW AND SIMULATED ANALYSES

History of Standardized Testing

The early history of standardized testing goes back several centuries. In the 3rd century BCE in imperial China, to qualify for civil service, Chinese aristocrats were examined for their proficiency in music, archery, horsemanship, calligraphy, arithmetic, and ceremonial knowledge. Later, the examinations tested knowledge of civil law, military affairs, agriculture, geography, composition, and poetry.[2,3] Those who passed these exams were qualified to serve the Chinese emperor and his family. The exams were accompanied by an atmosphere of solemnity and attention to the young nobles who dared to be scrutinized for the prestigious positions. The topics of the exams were frequently provided by the emperor, and he often examined the applicants during the final stage of the competition.

In the late 1880s, Francis Galton was inspired by the work of his cousin, Charles Darwin, regarding the origin of species and became interested in the hereditary basis of intelligence and the measurement of human ability. Galton developed the theoretical bases of testing—the application of a series of identical tests to a large number of individuals and the statistical processing of the results.[4] In 1904, Alfred Binet, a Parisian with a doctorate in experimental psychology, was commissioned by the French ministry of education to study schoolchildren who were developmentally behind their peers. His task was to develop a method to identify children who were not benefiting from inclusion in regular classrooms and required special education.[5] For this purpose, Binet and his associate, Theodore Simon, designed and administered a 30-item instrument arranged by difficulty that tested ability for judgment, understanding, and reasoning.[1]

The field of testing developed rapidly during World War I (1914–1918), when the problem of professional selection for the needs of the army and military production became a priority. During that time, leading psychologists organized the Army Alpha Examination to test army recruits.[6] Their success further inspired psychologists to advocate for civilian testing. During the 20th century, large-scale assessment in the United States became a necessity for college admissions and school accountability. The reliance on standardized tests for college admission was a response to the increasing number of students applying to colleges, and it became a tool to tighten the gates in the face of limited resources.[7]

In the 21st century, standardized tests constitute an inseparable part of American culture. Assessment instruments are administered in a wide range of settings: K–12, college admission, academic progression, professional licensure, clinical credentialing, industrial, forensic, and many more. "Gatekeepers of America's meritocracy—educators, academic institutions, and employers—have used test scores to label people as bright or not bright, as worthy academically or not worthy."[8] The study of measurement processes and the methods used to produce scores in testing evolved into a specialized discipline—*psychometrics*, a combination of education, psychology, and statistics.[9]

Critique of Standardized Tests

As the use of standardized tests for high-stakes exams increased, so did the critique of their use.[10] Counsell[11] conducted a case study exploring the effect of the high-stakes accountability system on the lives of students and teachers. The findings revealed that the culture of testing introduces a continuum of fear and ethical and moral dilemmas related to the pressure experienced by instructors when schools use test scores as a measure of accountability. Often, instructors decontextualize the material to the students with an intention to artificially inflate the test scores.[12] Such a phenomenon is known to researchers as "teaching to the test" and is often controlled for by psychometric procedures.[13]

Kohn[14] claimed that admission tests (such as the SAT and ACT) are "not very effective as predictors of future academic performance, even in the freshman year of college, much less as predictors of professional success." Zwick and Himelfarb[15] predicted 1st-year undergraduate grade-point average (FYGPA) in 34 colleges from high school GPA (HSGPA) and SAT scores using linear regression models. The average R^2 for these regression models was .226 (this coefficient indicates the amount of variance in the regression outcome explained by the linear combination of the predictors). However, in most of the models, the HSGPA was the predictor that accounted for the majority of variance. Zwick and Himelfarb stated, "The only substantial increase in R^2 values occurred when SAT scores are added to a prediction equation that included self-reported HSGPA."

Furthermore, the study highlighted the overprediction (the predicted outcomes were higher than actual) of FYGPA for African American and Latino students and the underprediction (the predicted outcomes were lower than actual) for Caucasian and Asian students when high school grades and SAT scores were used. Zwick and Himelfarb concluded that these errors in prediction were partially attributed to high school socioeconomic status—African American and Latino students are more likely than Caucasian students to attend high schools with fewer resources.

Measurement and Classification

Two processes are involved when a test is administered—measurement and classification. Measurement is the process of assigning numerical values to a phenomenon. This is a thorny process because numbers are used to categorize the phenomenon, and numerical scales hold qualities such as differentiation (1 is different from 2), order (2 is higher than 1), equality of intervals (the interval between 1 and 2 is equal to the interval between 2 and 3), and a 0 point, which is not always a true absence of value. By assigning numerical values to categories, the rules associated with numbers are carried over to the properties of the measured phenomenon and may not always correspond to the actual properties of the measured objects.

Stevens[16] developed a hierarchy of measurement scales: nominal, ordinal, interval, and ratio. The *nominal* scale is a system of measurement where numbers are used for the purpose of differentiation only. For example, the numerical part of a street address or apartment number is numbered on the nominal scale. The number on the jersey of a football player is used to differentiate the player from others, and it too is on the nominal scale. The categorical coding of most demographic variables, such as gender, ethnicity, and political party affiliation, constitutes nominal measures.[17] Since nominal enumeration is used only to distinguish categories, the numbers assigned to the categories do not follow any order or presume interval equality. The nominal scale is the most rudimentary form of measurement.

The *ordinal* scale is a measurement scheme where, in addition to simple differentiation (the attribute specified by the nominal scale), the numbers represent a rank order of the measured phenomenon. Examples of ordinal measures are rankings in the Olympic Games, progressions of the spiciness of a dish in a restaurant (mild, spicy, and very spicy), military rank, birth order, and class rank. Another example of an ordinal measure is the emoji-face pain scale commonly used in health care. An ordinal scale establishes the order of categories but lacks the ability of comparison between the categories' intervals.

The subsequent scale in Stevens's hierarchy is the *interval* scale, which, in addition to differentiation and rank order, establishes the property of interval equality. On this scale, the intervals between adjacent points are

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

presumed to be equal. One example of the interval scale is a number line, where, going from left to right, each subsequent number is higher in rank, and the intervals between adjacent numbers are equal across the entire domain of the line. Another example is a temperature scale measured in Celsius or Fahrenheit. In the social sciences, items commonly measured on the Likert scale, ranging from "strongly disagree" to "strongly agree," for the purposes of statistical analysis of opinions, are assumed to be on the interval scale.

The highest measurement scale in the hierarchy is the *ratio* scale. In addition to the properties established by the nominal, ordinal, and interval scales, a ratio scale has a true 0 point (complete absence of value). Neither the number line nor the Celsius or Fahrenheit temperature scales have an absolute 0 point. The 0 on the number line is nothing more than a separation between the negative and positive numbers and can be rescaled with a simple linear transformation. The 0 on the temperature scale (in Celsius) is also not an absence of value but rather a point at which water becomes ice. An example of a ratio scale is the Kelvin temperature scale, where 0 indicates a complete absence of temperature.

Every assessment is designed to measure and classify the test takers' performance in a specific domain. Depending on the assessment design, the scores can be on the ordinal, interval, or even ratio scale. Then, depending on the score obtained on the test, a test taker can be classified into the mastery or nonmastery categories (in the case of professional testing) or into basic, proficient, or advanced levels of performance in the case of K–12.[18]

When test takers present themselves at the test site for an exam administration, they arrive as members of a single population. The goal of the test designer and test administrator is to separate the test takers into subpopulations according to the intended users' objectives for the scores. Thus, each item on the test is a classification tool that helps make the categorization decision regarding each individual test taker. With each item that is answered correctly, a test taker is more likely to be classified into the higher category, while each incorrect response increases the likelihood of classification into a lower category.

Reliability and Validity

The quality of a measurement instrument is expressed in terms of the *reliability* and *validity* of the scores collected by this instrument. Reliability is the consistency with which a measure, scale, or instrument assesses a given construct, while validity refers to the degree of relationship, or the "overlap" between an instrument and the construct it is intended to measure.[13] The traditional meaning of reliability is the degree to which respondents' scores on a given administration of a measure resemble their scores on the same instrument administered later within a reasonable time frame. Kerlinger and Lee[19] suggested 3 approaches to reliability: stability, lack of distortion, and being free of measurement error. The first 2 definitions are addressed in this section; the third definition requires an introduction to classical test theory[20,21] and is addressed later.

If a measurement instrument or a comparable form is administered multiple times to the same or a similar group of people, we should expect similar scores. This is called *temporal stability*—the degree to which data obtained in a given test administration resemble those obtained in following administrations. When an assessment is conducted, a score user expects assurance that scores are replicable if the same individuals are tested repeatedly under the same circumstances.[9] There are 2 techniques to assess temporal stability: the test–retest method and the parallel forms method.

In the test–retest method, a set of items is administered to a group of subjects, then the test is readministered later to the same group. The correlation of the 2 sets of scores is then measured. A higher correlation between the scores indicates higher reliability.

In the parallel forms method, 2 different forms of the same test are constructed, both measuring the same critical trait (knowledge base). Next, both forms are administered to the same group of test takers at the same test session. A higher relationship between the 2 sets of scores indicates higher reliability. However, it is very difficult to correctly construct equivalent test forms, and a weak relationship between the 2 sets of scores may actually reflect a lack of equivalence.

Another component of reliability is a scale's *internal consistency*. The lack of distortion or internal consistency of an instrument refers to the extent to which the individual components of a test are interrelated and thus produce the same or similar results. Items on the test should "hang together." One of the earlier techniques to establish the internal consistency of a scale is known as the split-half reliability.[22] The test is randomly split in half, and the 2 sets of test scores are compared to each other. Once again, a closer relationship between the 2 sets of scores indicates a higher test reliability.

Cronbach[6,23] developed the *coefficient alpha*, an alternative to the once common split-half technique, which has become the most universal technique for estimating internal consistency reliability. His coefficient alpha assesses reliability as a ratio of the summed variances of individual items and the total variance for the instrument, subtracted from 1 and adjusted for the number of items in the instrument. Cronbach's alpha coefficient is computed as follows:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_T^2}\right) \qquad (1)$$

where α is the estimate of the instrument's internal consistency reliability; k is the number of items on the instrument; i is the item indicator, $i = 1, 2, \ldots, k$; σ_i^2 is the variance of item i; and σ_T^2 is the total variance of the scale.

Cronbach's alpha ranges from 0 to 1.0 with values closer to 1.0 indicating higher reliability. The internal consistency of a test is considered acceptable if the alpha coefficient is above .70.[24,25] An alternative interpretation of Cronbach's alpha is the mean of all interitem correlations. If a correlation coefficient is squared, it becomes a *coefficient of determination*, which indicates the proportion of vari-

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

ability shared between 2 variables.[19] Thus, when .70 is squared, it becomes .49. This means that at least half of the variability in the responses collected by the instrument is explained by the instrument's internal consistency.

Reliability alone is not sufficient to establish the quality of a test. A good test must also measure what it was designed to measure, which is often referred to as *validity*. The validity of a scale refers to the extent of correspondence between variations in the scores on the test and the variation among respondents on the underlying construct being tested.[13] The process of validation is closely related to the intended use of the scores. For example, scores collected on a test of general anatomy given in English ideally depict the knowledge of anatomy possessed by a test taker. Yet, if a test is given to a sample of English-language learners, a part of the variability in scores can be explained by English proficiency (or lack thereof). Therefore, the scores collected by the same test in an English-first population of test takers may have higher validity than scores collected from English-language learners.

Importantly, the validity of a test is a matter of degree, not all or none. Further, the existing evidence of validity may be challenged by new findings or by new circumstances. Unavoidably, validity becomes an evolving property, and test validation is a continuous process.[26] This process of validation requires ongoing empirical research efforts outside of those used for reliability. The methods employed for establishing validity of a test include a thorough analysis of the content of the test during the phase of its scale development and quantitative assessment of the relationship between the test scores and the criterion that has been tested.[2] The degree of accuracy with which test scores relate to their intended use may be established by studying the predictive validity.

Test scores with low validity can still be reliable, while reliability is a prerequisite for validity. Establishing reliability is more of a technical matter, whereas validity requires much deeper thinking and consideration; it is much more than a statistical procedure. Continuous vigilant consideration of each item in terms of content representation and its statistical performance as well as the reflection on the populations of test takers are all essential for confirming a test score's validity.

Classical Test Theory

Any measurement is an inference, and any statistical inference is subject to error. All measurements are susceptible to random error and, if repeated, may vary. To comprehend the size and the origin of the error, ideally, the measurement should be repeated several times, as the average of a series of measurements is more precise than any individual measurement by a factor equal to the square root of the number of measurements.[27] Classical test theory (CTT) postulates that any observation is a linear combination of the true score and error. The fundamental equation of CTT states the following:

$$O_i = T_i + E_i \tag{2}$$

where O_i is the observed score for an examinee *i*, T_i is the true score for that examinee, and E_i is the error in the measurement. Thus, every test could be seen as a combination of 2 hypothetical components: the true score (true knowledge of the material tested) and the deviations from the true score due to random or systematic factors. Any systematic errors in measurement become part of an individual's true score and affect the validity since the score is no longer an estimate only of the latent trait but also of the systematic variability. The random errors, on the other hand, affect the reliability of the score and create a distortion in the observed score's precision over repeated administrations of the test.

Test scores can be described as random variables.[9] A random variable X is an outcome of a process that is determined by a probability distribution. The term "expectation" or "expected value," denoted as $E(X)$, is used to signify the mean of the probability distribution. Assuming that all systematic variability in the observed score is accounted for by the true score and the error component consists of only random error, we can specify the distribution of the errors as follows:

$$e_i \sim N(0, \sigma^2) \tag{3}$$

which means that if examinee *i* takes the exam an infinite number of times, by definition of random, the same amount of error will be distributed above and below the true score. Thus, the error will average at 0. The relationship between the observed score and the true score can be clarified by taking the expectation of the observed score:

$$E(O_i) = E(T_i) + E(E_i) \tag{4}$$

Meanwhile, if the expectation of error is 0 (see equation 3) and the expected value of the observed score is the true score,

$$E(O_i) = E(T_i) + 0 = E(T_i) = T_i \tag{5}$$

Then it follows from equations 2 and 5 that

$$E(E_i) = E(O_i) - E(T_i) = T_i - T_i = 0 \tag{6}$$

There are 3 other fundamental assumptions made by CTT: it is assumed that the correlation between true score and error is 0, that the correlation between error score on test 1 and error score on test 2 is 0, and that the correlation between the true score on test 1 and the error score on test 2 is 0.

The definition of reliability can be formulated in the framework of CTT if the following extension is made to the equation 2:

$$Var(O_i) = Var(T_i) + Var(E_i) \tag{7}$$

where $Var(O_i)$, the observed score variability, is partitioned into the true score variability, $Var(T_i)$, and the variability of error, $Var(E_i)$. Reliability is the proportion of the true score variability to the observed score variability or the proportion of the error variability to the observed score variability subtracted from 1.0:

$$\rho_{O1,O2} = \frac{Var(T_i)}{Var(O_i)} = 1 - \frac{Var(E_i)}{Var(O_i)} \tag{8}$$

with $\rho_{O1,O2}$ being the reliability coefficient.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

The variability of the scores, as viewed by CTT, provides the explanation for score stability. Test takers who are not satisfied with their exam scores may choose to repeat the test. While an examinee repeating a test is interested in the increase of the observed score, psychometricians consider any increase in the true score separately from the increase in the error component. If a test is reliable, it is very hard to increase the true score component when the assessment is repeated over a short period of time. Only long-term learning is associated with an increase in the true score component.[28,29] At the same time, the scores for a repeat test taker will vary from 1 administration to another, and, usually, improved performance may be seen on a second measurement occasion, even if different questions are used.[12] This is due to the known phenomenon called the *practice effect*,[30] which is defined as an increase in an examinee's test score from 1 administration of the same assessment to the next in the absence of learning, coaching, or other factors that are known to increase the score.[31]

Other sources of measurement error may include temporary or momentary fatigue, fluctuations of memory or mood, or fortuitous conditions at a particular time that temporarily affect the outcomes measured by the test.[19] Test scores may also be influenced by the content of the material that appeared on the test, guessing, state of alertness, and even scoring errors.

Another likely explanation of the differences in scores from 1 measurement occasion to another is the phenomenon known as *regression to the mean*.[32] Each form of a test will tend to favor certain students but not others in a nonsystematic way. Students may get a test with items representing the material they are most familiar with or have studied the most. However, students who were favored by 1 form of the test are not likely to be favored by another when they retake the test. Therefore, the scores obtained on the second or third testing occasions will tend to be closer to the mean than the scores obtained on the first testing occasion.[33]

Even though it is never possible to measure exactly how much an increase in the observed score is influenced by the error component, CTT allows for estimation of the *standard error of measurement* (SEM), which is a function of the standard deviation of the set of observed scores and the reliability of the test:

$$SEM = SD_O\sqrt{1 - \hat{\rho}_{O1,O2}} \tag{9}$$

where SD_O is the standard deviation of the set of observed scores and $\hat{\rho}_{O1,O2}$ is an estimate of reliability. Estimates of the SEM can be helpful in interpreting increases in individual test scores.

Item Response Theory

Item response theory (IRT) is a collection of statistical and psychometric methods used to model test takers' item responses.[34] The initial development of IRT models took place in the second half of the 20th century. First, Rasch[35] developed a model for analyzing categorical data. Next, Lord and Novick[21] wrote chapters on the theory of latent trait estimation, which gave birth to a new way of data analysis in testing. Prior to the development of IRT, the testing industry relied on CTT methods for modeling test item responses. Since then, IRT has made its way into every aspect of the testing industry. IRT methods are used today in test development, item banking, data analysis, analysis of differential item functioning, adaptive testing, test equating, and test scaling.[36]

The early IRT models were first developed for dichotomously scored item responses (eg, 0 = wrong, 1 = right). These models included the 1-parameter logistic model (1PL), the 2-parameter logistic model (2PL), and the 3-parameter logistic model (3PL). Common assumptions for the early IRT models include *unidimensionality*—only 1 latent trait is necessary to explain the pattern of item-level responses[37]—and *local independence*—after accounting for the latent trait, there is no dependency among the items.[36] Later, models for polytomous responses were developed: the partial credit model[38] and the generalized partial credit model.[35]

In the early 1990s, significant efforts were made to develop multidimensional IRT models[39,40] and models that were able to account for item dependency over and above the dependency explained by the common trait.[41,42] Due to the introductory nature of this article, I will present the mathematical logic and graphical examples of the 1PL, 2PL, and 3PL models only.

One advantage of IRT over traditional testing theories is that IRT defines a scale for the underlying latent variable that is being measured by the test items.[43] IRT assumes that responses on a unidimensional test are underlined by a single latent trait (θ), often called the test taker's "ability." This latent trait is not able to be observed directly; however, it can be constructed using observed responses to the items on a test. Assuming IRT, the probability of a response to an item on a test is conditional on θ:

$$f_i(u_i|\ \theta) = P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} \tag{10}$$

where $f_i(u_i|\theta)$ is the function of providing response u on an item i conditional on ability θ, $P_i(\theta)^{u_i}$ is the probability of a correct response ($u_i = 1$), and $Q_i(\theta)^{1-u_i}$ is the probability of an incorrect response ($u_i = 0$) Subsequently, if ($u_i = 1$), $f_i(u_i|\theta) = Pi(\theta)$, and if $u_i = 0$, then $f_i(u_i|\theta) = Q_i(\theta)$. The function connecting the means of conditional distributions (equation 10) is the regression of the item score on ability and is referred to as the "item characteristic curve" (ICC). The ICC relates the probability of providing a correct response on an item to the ability measured by the entire test.[37]

The student's ability and the item difficulty are on the same scale; therefore, $\theta_j = \beta_i$ corresponds to $\theta - \beta = 0$, meaning that there is an exact match between an examinee's ability and item difficulty; $\theta_j > \beta_i$ corresponds to $\theta - \beta > 0$, which means that the item is easy for the examinee's ability level; and $\theta_j < \beta_i$ means that when $\theta - \beta < 0$, the item is difficult for the test taker. Thus, the probability of providing a correct response by an examinee j to an item i is a function of the difference between theta and beta; formulaically,

$$P(u_i = 1|\ \theta, \beta) = f(\theta_j - \beta_{i,}) \tag{11}$$

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

where f is a function that relates the ability and the probability (ICC).

1PL Model

In this model, the probability of the response to an item is a function of the difference between the test taker's ability and the item's difficulty. The following is the equation for 1PL:

$$P(u_i = 1|\theta, \beta) = \frac{e^{D(\theta_j - \beta_i)}}{1 + e^{D(\theta_j - \beta_i)}} \tag{12}$$

where D is a scaling factor, set to $D = 1.7$, so the values of $P(\theta)$ for 2-parameter normal ogive and the values for 2PL differ by less than 0.01.

Illustration

The computing language R (an open-source environment for statistical computing and graphics) is often used to fit IRT models to data and estimate item parameters. Presented here is an example by means of the "irtoys" package[44] to fit various IRT models using a set of simulated responses (n = 2000) to a 40-item test. The items were scored dichotomously. Table 1 presents estimates of model parameters and associated standard errors for the 1PL model. The item difficulty is the only parameter that was estimated, while the item discrimination was fixed at 1. Figure 1a presents the ICC curves for the 40 items. The curves differ by their location in relation to the x-axis, which is a reference scale for the test takers' ability and item difficulty—more difficult items are to the right, while less difficult items are to the left. The 1PL model assumes that all items relate to the latent trait (ability) equally and differ only in the amount of difficulty.

Figure 1b presents the item information functions (IIF) for the 40 items. The IIF shows the point on the ability scale for which the item provides maximum information. Assuming that these curves are Gaussian, the ranges of ability for which an item provides the most information can be estimated using the 3-sigma empirical rule.[45] The IIF depends on the slope of the item response function as well as the conditional variance at each ability level. The greater the slope and the smaller the variance, the greater the information and the smaller the standard error of measurement (SEM).[32] In 1PL, the slopes are held constant; therefore, there is no variability in the height of the curves.

2PL Model

The 2PL model estimates another parameter—the *discrimination* of an item, seen as the slope of the ICC. The discrimination is between those test takers who know the right answer and the population of test takers who do not demonstrate that knowledge. The items with better discriminating qualities have steeper slopes. The following equation represents the 2PL model:

$$P(u_i = 1|\theta, \alpha, \beta) = \frac{e^{Da_i(\theta_j - \beta_i)}}{1 + e^{Da_i(\theta_j - \beta_i)}} \tag{13}$$

where a_i is the discrimination parameter for item i. Table 2 presents the model parameter estimates and related standard errors for the 2PL model. Figure 2a presents the ICCs for the same 40 items as Figure 1a; it is now obvious that some items are better at discriminating between the 2 populations (have steeper slopes) than others.

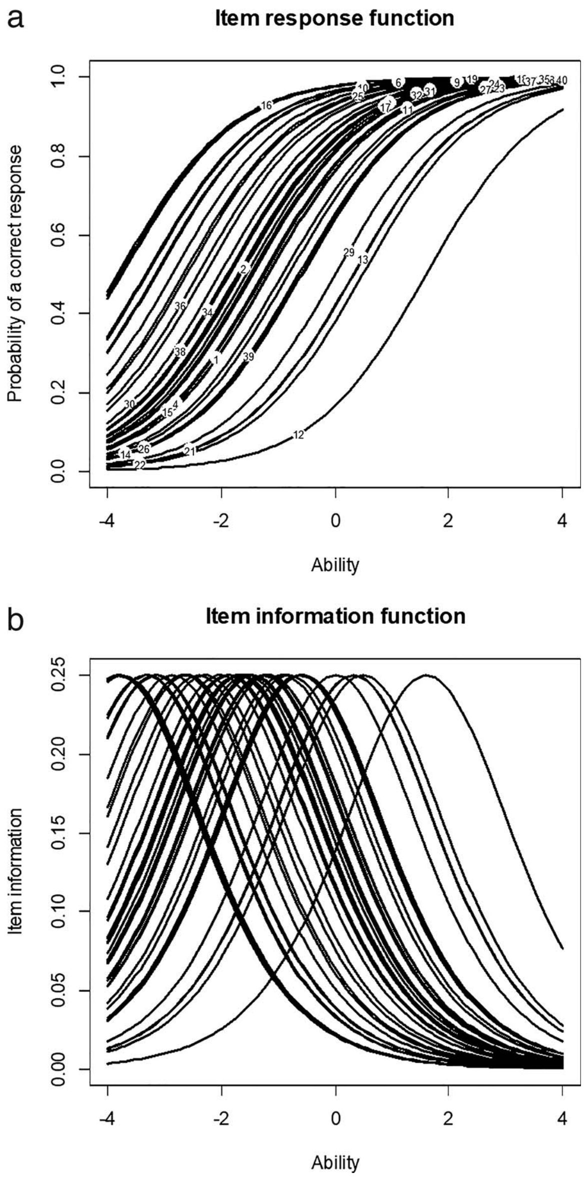

Figure 1 - a) Item characteristic curves for the 40 items, 1-parameter logistic model. b) Item information functions for the 40 items, 1-parameter logistic model.

The estimation of the slope relaxes the assumption of an invariant relationship between the items and the latent trait. This relationship can now be estimated, and it is similar to the factor loadings in factor analysis.[46] The items with higher discrimination coefficients are more responsive to small changes in the latent trait, whereas the items with low discrimination coefficients require large changes in the latent trait to reflect a change in the probability. Figure 2b

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

Table 1 - Item-Parameter Estimates, 1-Parameter Logistic Model (N/A = Not Applicable)

Item	α	β	γ	SE α	SE β	SE γ
1	1	−1.17	N/A	N/A	0.08	N/A
2	1	−1.66	N/A	N/A	0.09	N/A
3	1	−1.71	N/A	N/A	0.10	N/A
4	1	−1.24	N/A	N/A	0.09	N/A
5	1	−2.87	N/A	N/A	0.14	N/A
6	1	−3.34	N/A	N/A	0.17	N/A
7	1	−3.78	N/A	N/A	0.20	N/A
8	1	−3.32	N/A	N/A	0.16	N/A
9	1	−2.30	N/A	N/A	0.11	N/A
10	1	−3.15	N/A	N/A	0.15	N/A
11	1	−1.18	N/A	N/A	0.09	N/A
12	1	1.60	N/A	N/A	0.09	N/A
13	1	0.31	N/A	N/A	0.08	N/A
14	1	−0.60	N/A	N/A	0.08	N/A
15	1	−1.26	N/A	N/A	0.09	N/A
16	1	−3.82	N/A	N/A	0.20	N/A
17	1	−1.67	N/A	N/A	0.09	N/A
18	1	−3.17	N/A	N/A	0.15	N/A
19	1	−3.75	N/A	N/A	0.20	N/A
20	1	−1.67	N/A	N/A	0.09	N/A
21	1	0.32	N/A	N/A	0.08	N/A
22	1	0.48	N/A	N/A	0.08	N/A
23	1	−0.82	N/A	N/A	0.08	N/A
24	1	−1.49	N/A	N/A	0.09	N/A
25	1	−2.68	N/A	N/A	0.13	N/A
26	1	−0.58	N/A	N/A	0.08	N/A
27	1	−0.92	N/A	N/A	0.08	N/A
28	1	−1.56	N/A	N/A	0.09	N/A
29	1	0.00	N/A	N/A	0.08	N/A
30	1	−2.03	N/A	N/A	0.10	N/A
31	1	−1.67	N/A	N/A	0.09	N/A
32	1	−1.70	N/A	N/A	0.09	N/A
33	1	−1.27	N/A	N/A	0.09	N/A
34	1	−1.86	N/A	N/A	0.10	N/A
35	1	−2.62	N/A	N/A	0.12	N/A
36	1	−2.42	N/A	N/A	0.12	N/A
37	1	−1.45	N/A	N/A	0.09	N/A
38	1	−1.90	N/A	N/A	0.10	N/A
39	1	−0.64	N/A	N/A	0.08	N/A
40	1	−1.47	N/A	N/A	0.09	N/A

presents the items' information curves, which now show variability in the amount of information they provide.

3PL Model

The 3PL model is a 2PL model with an additional parameter, γ_i, which is the lower asymptote of the ICC and represents the probability of a test taker with a low ability providing a correct answer to an item *i*. The inclusion of this parameter suggests that test takers who score low on the latent trait may still provide a correct response by chance. This parameter is referred to as "guessing." The following is the mathematical representation of the 3PL model:

$$P(u_i = 1|\theta, \alpha, \beta, \gamma) = \gamma_i + (1 - \gamma_i)\frac{e^{Da_i(\theta_j - \beta_i)}}{1 + e^{Da_i(\theta_j - \beta_i)}} \qquad (14)$$

where γ_i is the guessing parameter. Referring back to equation 14, if a test taker guessed ($\gamma_i = 1$), then the probability of the correct response is entirely explained by guessing (the term after the plus sign disappears). However, if the test taker did not guess ($\gamma_i = 0$), the model defaults to the 2PL. Table 3 presents model parameter estimates for the 3PL, while Figure 3a and b presents ICCs and IIFs, respectively, for the 40 items.

Polytomous IRT Models

Various polytomous IRT models have been developed to account for ordered categorical responses. Samejima[47] developed a logistic model for graded responses in which the probability that an examinee *j* with a particular level of ability will provide a response to an item *i* of the category *k* is the difference between the cumulative probability of a

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

Table 2 - Item-Parameter Estimates, 2-Parameter Logistic Model (N/A = Not Applicable)

Item	α	β	γ	SE α	SE β	SE γ
1	0.39	−2.77	N/A	0.09	0.67	N/A
2	0.70	−2.33	N/A	0.12	0.35	N/A
3	0.62	−2.66	N/A	0.11	0.45	N/A
4	0.81	−1.55	N/A	0.11	0.20	N/A
5	0.84	−3.47	N/A	0.18	0.62	N/A
6	1.01	−3.48	N/A	0.22	0.60	N/A
7	0.84	−4.57	N/A	0.25	1.17	N/A
8	1.35	−2.81	N/A	0.24	0.36	N/A
9	0.89	−2.65	N/A	0.15	0.37	N/A
10	1.39	−2.64	N/A	0.24	0.31	N/A
11	1.00	−1.26	N/A	0.12	0.14	N/A
12	0.45	3.30	N/A	0.11	0.74	N/A
13	0.88	0.37	N/A	0.11	0.09	N/A
14	0.78	−0.77	N/A	0.10	0.13	N/A
15	0.50	−2.36	N/A	0.10	0.46	N/A
16	0.64	−5.82	N/A	0.25	2.04	N/A
17	0.41	−3.75	N/A	0.11	0.93	N/A
18	0.87	−3.71	N/A	0.20	0.70	N/A
19	1.06	−3.76	N/A	0.26	0.73	N/A
20	0.77	−2.16	N/A	0.12	0.30	N/A
21	0.79	0.41	N/A	0.10	0.10	N/A
22	0.12	3.48	N/A	0.08	2.32	N/A
23	0.49	−1.57	N/A	0.09	0.31	N/A
24	0.57	−2.49	N/A	0.11	0.44	N/A
25	0.70	−3.75	N/A	0.16	0.75	N/A
26	0.75	−0.77	N/A	0.10	0.13	N/A
27	0.16	−5.10	N/A	0.09	2.73	N/A
28	1.08	−1.57	N/A	0.14	0.16	N/A
29	0.25	0.02	N/A	0.08	0.26	N/A
30	0.33	−5.66	N/A	0.12	1.94	N/A
31	0.92	−1.89	N/A	0.13	0.23	N/A
32	0.32	−4.90	N/A	0.11	1.58	N/A
33	0.66	−1.88	N/A	0.11	0.28	N/A
34	0.76	−2.45	N/A	0.13	0.35	N/A
35	1.64	−2.01	N/A	0.23	0.18	N/A
36	0.53	−4.33	N/A	0.14	1.04	N/A
37	0.81	−1.80	N/A	0.12	0.23	N/A
38	0.52	−3.45	N/A	0.12	0.72	N/A
39	0.39	−1.49	N/A	0.09	0.36	N/A
40	0.62	−2.26	N/A	0.11	0.36	N/A

response to that category or higher and the cumulative probability of a response to the next highest category or higher. Consider the following:

$$P_{ijk}(\theta)_j = P'_{ijk}(\theta)_j - P'_{ijk+1}(\theta)_j$$

$$P'_{ijk}(\theta)_j = \frac{1}{1 + \exp[-Da_i(\theta_j - b_{ik})]}$$

where b_{ik} is the difficulty parameter for category k_i and a_i is the discrimination parameter for item j.[47]

A different model for ordered categorical response was developed by Masters.[33] In this partial credit model, the probability that an examinee j will provide a response x on item i with M_i thresholds is a function of student's ability and the difficulties from the M_i thresholds in item i is given by the following:

$$P_{ijx} = \frac{exp\sum_{x=0}^{X}(\theta_j - b_{ix})}{\sum_{m=0}^{M}(exp\sum_{x=0}^{m}(\theta_j - b_{ix}))}$$

where $x = 1,2, \ldots, M_i$ is the count of successfully completed thresholds, and $\sum(\theta_j - b_{ix}) = 0$.[33]

Samejima's graded response model was fitted to a simulated data set of $n = 1500$ responses to 10 polytomous items scored using the following categories: 0, 1, 2, and 3. Table 4 presents model-based parameter estimates; Figure 4a presents ICC curves for items 1–4 of the 10 polytomous items. Figure 4b and c presents ICC curves for items 5–8 and 9 and 10, respectively.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

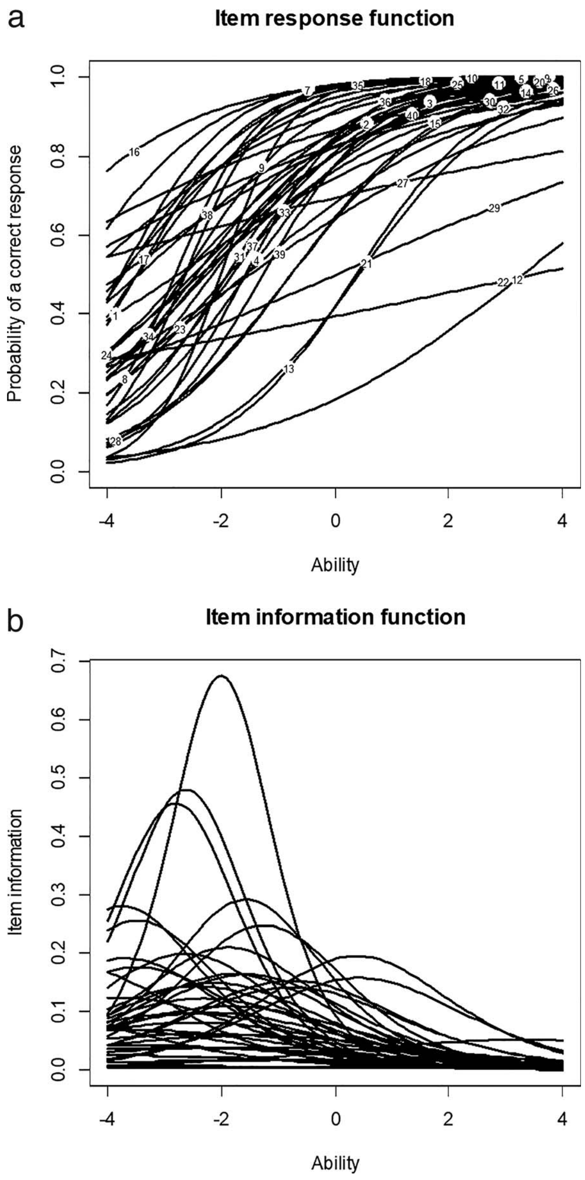

Figure 2 - a) Item characteristic curves for the 40 items, 2-parameter logistic model. b) Item information functions for the 40 items, 2-parameter logistic model.

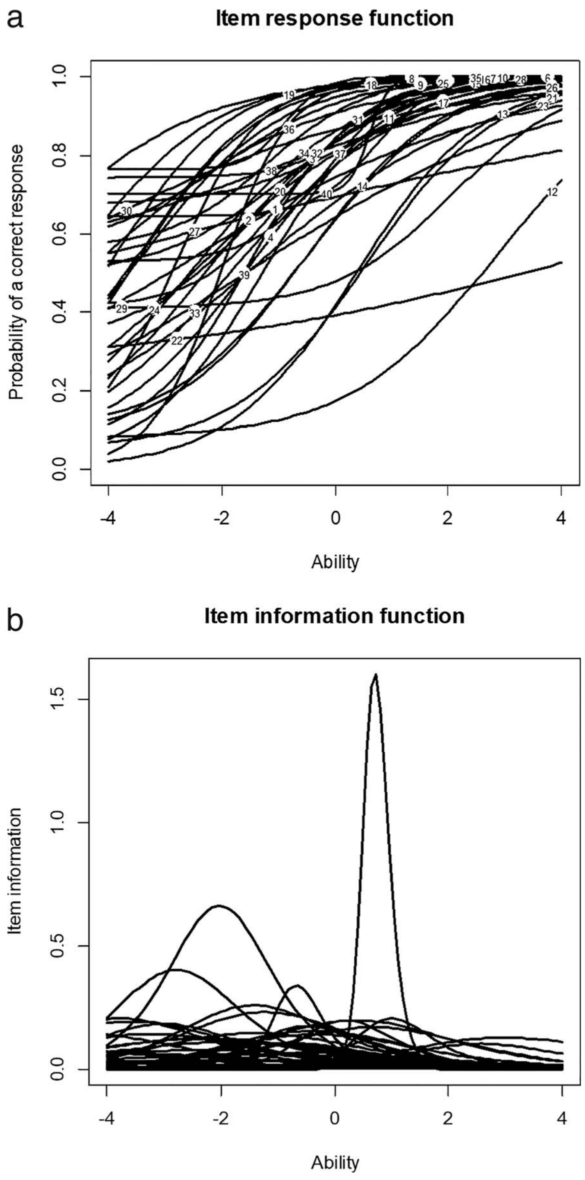

Figure 3 - a) Item characteristic curves for the 40 items, 3-parameter logistic model. b) Item information functions for the 40 items, 3-parameter logistic model.

Equating

Measurements of the same construct collected at different times or by different forms must be brought to the same scale to be comparable. In the field of testing, when tests are used to make high-stakes decisions, the scores for examinees who took the test on 1 occasion using 1 test form should be comparable to the scores of examinees who took the test on another occasion using a different test form. Due to the security of test programs, it is common practice to administer different forms of the test on different testing occasions. However, it is hard to construct 2 truly parallel forms, and often these test forms differ in difficulty. Yet it is important to avoid a situation where 1 group of test takers has an unfair advantage because they were administered an easier form of the exam.[48] Therefore, the test scores must be equated to account for the possible differences in difficulty between the test forms or differences in ability between the groups of test takers.

Equating is a statistical process used to adjust scores on test forms so that scores on the forms can be used interchangeably.[36] After equating, alternate forms of the same test yield scaled scores that can be used interchangeably even though they are based on different sets of items.[49] It is important to point out that statistical adjustment is not possible for differences in content. The responsibility for the content equivalence between 2 forms of a test lies entirely on test developers.

For the past 30 years, equating has received much deserved attention and research. Many new equating

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

Table 3 - Item-Parameter Estimates, 3-Parameter Logistic Model (N/A = Not Applicable)

Item	α	β	γ	SE α	SE β	SE γ
1	0.37	−2.68	0.05	0.97	0.46	0.04
2	0.71	−2.30	0.00	0.17	0.78	0.69
3	0.86	−0.84	0.49	0.71	0.76	0.11
4	0.79	−1.50	0.04	0.35	0.86	0.53
5	0.82	−3.42	0.09	0.11	0.69	0.10
6	1.01	−3.28	0.17	0.13	0.44	0.10
7	0.81	−4.60	0.07	0.37	0.11	0.47
8	1.32	−2.84	0.04	0.84	0.48	0.11
9	1.14	−1.18	0.56	0.37	0.83	0.31
10	2.98	−0.88	0.76	0.38	0.20	0.08
11	0.97	−1.26	0.01	0.15	0.41	0.17
12	0.77	2.82	0.08	0.41	0.15	1.06
13	0.89	0.37	0.00	0.43	0.82	0.24
14	0.85	−0.53	0.08	0.14	0.94	0.73
15	1.84	0.70	0.64	0.13	0.82	0.14
16	0.64	−5.75	0.06	2.20	0.34	0.04
17	0.92	0.34	0.67	1.00	0.29	0.06
18	0.90	−3.59	0.03	0.42	0.58	0.19
19	0.98	−3.89	0.09	0.39	0.50	0.15
20	0.88	−1.33	0.31	0.17	0.53	0.07
21	0.87	0.56	0.05	0.19	0.54	0.09
22	0.22	6.28	0.24	0.46	0.30	0.10
23	0.49	−1.48	0.02	0.12	0.66	0.35
24	0.57	−2.42	0.03	0.30	0.63	0.18
25	0.86	−2.02	0.55	0.41	0.59	0.13
26	0.82	−0.49	0.09	0.51	0.42	0.11
27	0.17	−3.94	0.10	0.09	0.60	0.10
28	1.07	−1.51	0.05	0.10	1.16	0.12
29	0.96	2.10	0.41	0.10	0.51	0.03
30	0.33	−5.40	0.07	0.11	0.10	0.02
31	1.25	−0.82	0.41	0.13	0.65	0.11
32	0.44	−1.28	0.52	0.49	0.53	0.15
33	0.67	−1.85	0.00	0.30	1.04	0.38
34	0.72	−2.48	0.04	0.35	0.44	0.15
35	1.63	−2.03	0.00	0.09	0.15	0.03
36	0.52	−4.34	0.02	0.45	0.68	0.14
37	1.54	−0.15	0.53	1.45	0.26	0.04
38	1.69	0.37	0.74	0.12	0.15	0.01
39	0.37	−1.54	0.01	0.71	0.95	0.34
40	5.72	0.60	0.70	1.16	0.42	0.06

Table 4 - Item-Parameter Estimates, Graded Response

Item	b1	b2	b3	a
1	−0.97	−0.56	−0.13	5.51
2	−2.64	−1.84	−0.55	2.00
3	−2.94	−1.22	0.26	1.23
4	−1.27	−0.6	0.68	0.77
5	−3.77	−1.98	0.25	0.93
6	−1.61	−0.57	0.46	1.23
7	−2.14	−0.87	0.81	1.39
8	−1.96	−0.48	1.22	1.19
9	−0.62	1.75	2.94	0.66
10	−1.59	−0.47	1.07	1.76

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

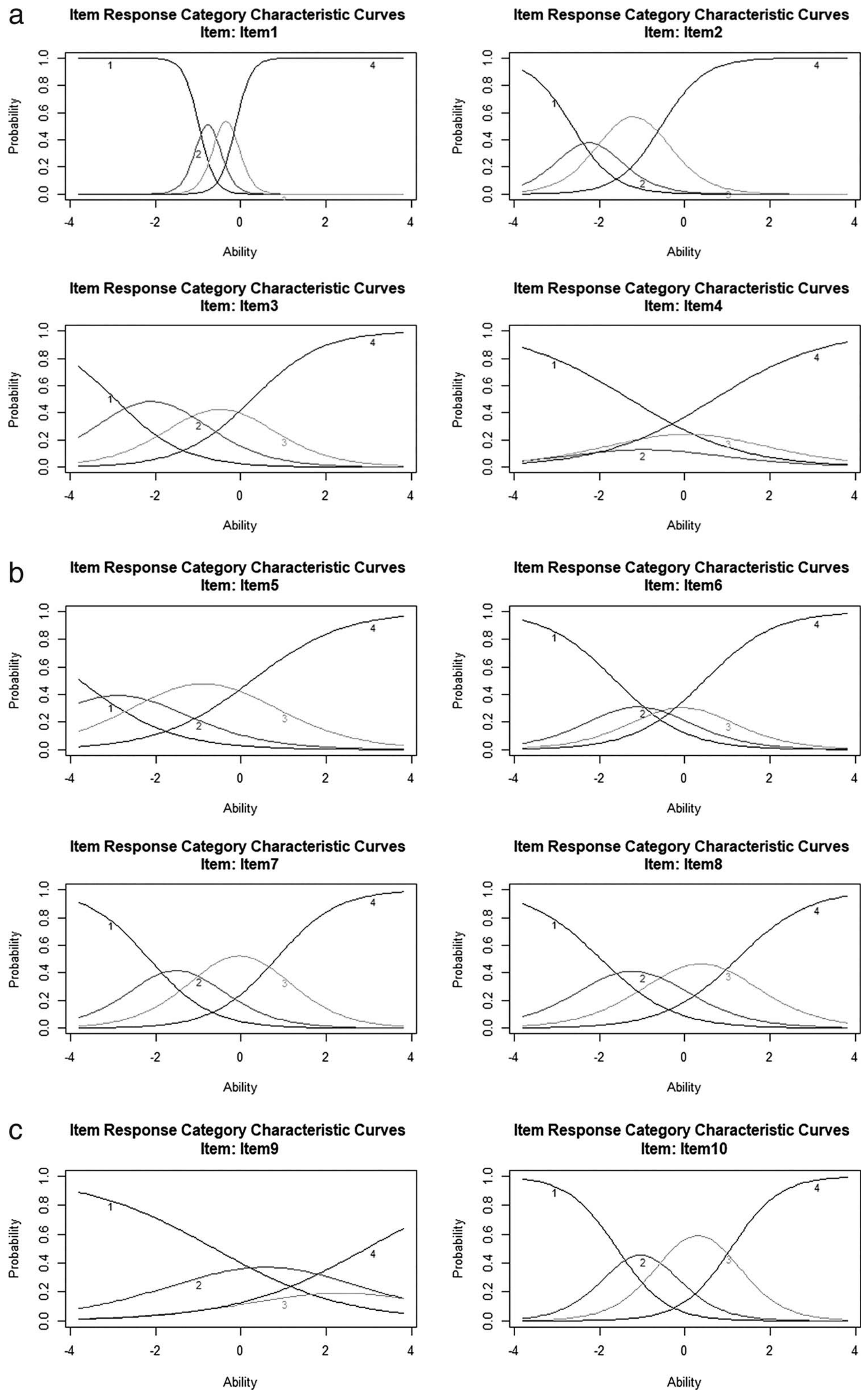

Figure 4 - a) Item characteristic curves for items 1-4, graded response. b) Item characteristic curves for items 5-8, graded response. c) Item characteristic curves for items 9 and 10, graded response.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

methods have been proposed and tested in both research and operational testing programs. I will introduce only general principles related to equating here, as my goal is to make the reader aware of the procedure. Those who wish to expand their knowledge of equating should turn to the literature published in the field of educational measurement.

The first step in the process of equating is to decide on an equating design. Test scores can be equated using either the same populations or the same items. *Single-group design* assumes that 2 test forms can be equated if they are given to the same population of examinees. Since the same examinees take both tests, the difficulty levels are not confounded by the ability of the examinees.[37] *Equivalent-group design* assumes that 2 test forms are given to similar but not the same populations of examinees. Reasonable group equivalence may be achieved through random assignment.[13]

Common-item design requires that both forms of the test contain a set of the same items, usually called "anchor" items; the forms are then administered to different populations of examinees. Subsequently, a function that relates the statistics computed for each anchor set will account for the differences in difficulty. This mathematical function is then used to equate the nonanchor items on both forms.[36,37]

An appropriate equating methodology must be chosen, depending on which theoretical framework is preferred by the testing program, to obtain the test-taker statistics and the item-level statistics. Equating methods have been developed based on both CTT and IRT. When pairs of statistical values for 2 forms have been obtained, a decision is made regarding the methods to be used to relate these exams. Several methods can be selected from the framework of linear modes for this; they include regression methods, mean and sigma procedures, or characteristic curves methods.

Linking

Equating is the strongest form of linking. The tests can be similar or even equivalent in content and different in difficulty, or they can be different in content and also in difficulty. When tests are different in content, the scores obtained on these exams may still need to be put on the same scale. In this case, the statistical process of adjusting the scores for difficulty is called linking. When linking is used for equating, the relationship is invariant across different populations.[36] The term *equating* is reserved for the situation when scores from 2 tests of the same content are linked. The statistical procedures used in equating may not differ for linking; however, no linking procedures can adjust for differences in content.

Conclusion

This article presents researchers and clinicians in the health sciences with an introduction to educational measurement—the history, theoretical frameworks of the CTT and IRT, and the most common IRT models used in modern testing.

ACKNOWLEDGMENTS

This article is dedicated to Dr Howard B. Lee, a mentor and friend.

FUNDING AND CONFLICTS OF INTEREST

No funding was received for this work, and the author has no conflicts of interest to declare relevant to this work.

About the Author

Igor Himelfarb is the director of the Department of Psychometrics and Research for the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; ihimelfarb@nbce.org). Address correspondence to Igor Himelfarb, National Board of Chiropractic Examiners, 901 54th Avenue, Greeley, CO 80634; ihimelfarb@nbce.org.

This article was received July 2, 2018; revised September 16, 2018, and December 20, 2018; and accepted December 27, 2018.

Author Contributions

Concept development: IH. Design: IH. Supervision: IH. Data collection/processing: IH. Analysis/interpretation: IH. Literature search: IH. Writing: IH. Critical review: IH.

© 2019 Association of Chiropractic Colleges

REFERENCES

1. Shmitt K. What is licensure? In: *Licensure Testing: Purposes, Procedures and Practices*. Lincoln: Buros Institute of Mental Measurement and University of Nebraska, Lincoln; 1995:1–32.
2. Aiken LR. *Psychological Testing and Assessment*. 11th ed. Needham Heights, MA: Pearson; 2003.
3. Zwick R. *Fair Game? The Use of Standardized Admissions Tests in Higher Education*. New York, NY: Routledge-Falmer; 2002.
4. McIntire SA, Miller LA. *Foundation of Psychological Testing: A Practical Approach*. Thousand Oaks, CA: Sage; 2007.
5. Geisinger KF, Usher-Tate BJ. Brief history of educational testing and psychometrics. In: Wells GS, Faulkner-Bond M, eds. *Educational Measurement from Foundation to the Future*. New York, NY: Guilford; 2016.
6. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297–334.
7. Zwick R. *Who Gets In? Strategies for Fair and Effective College Admissions*. Cambridge, MA: Harvard University Press; 2017.
8. Sacks P. *Standardized Minds*. Cambridge, MA: Harvard University Press; 1999.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access

9. Cronbach LJ. Beyond the two disciplines of scientific psychology. *Am Psychol*. 1975;(30):380–385.
10. Huddleston AP, Rockwell EC. Assessment for the masses: a historical critique of high-stakes testing in reading. *Texas J Literacy Educ*. 2015;3(1):38–49.
11. Counsell S. *What Happens When Veteran and Beginner Teachers' Life Histories Intersect With High-stakes Testing and What Does It Mean for Learners and Teaching Practice: The Making of a Culture of Fear* [dissertation]. Cedar Falls: University of Northern Iowa; 2007.
12. Welsh M, Eastwood M, D'Agostino JV. Conceptualizing teaching to the test under standards-based reform. *Appl Meas Educ*. 2014;27:98–114.
13. Crano WD, Brewer MB, Lac A. *Principles and Methods of Social Research*. 3rd ed. New York, NY: Routledge; 2014.
14. Kohn A. *The Case Against Standardized Testing: Raising the Scores, Running the Schools*. Portsmouth, NH: Heinemann; 2000.
15. Zwick R, Himelfarb I. The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *J Educ Meas*. 2011;48(2):101–121.
16. Stevens SS. On the theory of scales measurement. *Science*. 1946;103(2684):677–680.
17. Bandalos DL. *Measurement Theory and Applications for the Social Sciences*. New York, NY: Guilford; 2018.
18. National Assessment of Educational Progress. The national report card. https://nces.ed.gov/nationsreportcard/pdf/main2009/2010458.pdf. Published 2009. Accessed June 1, 2018.
19. Kerlinger FN, Lee HB. *Foundation of Behavioral Research*. 4th ed. New York, NY: Harcourt College Publishers; 2000.
20. Guiliksen H. *Theory of Mental Tests*. New York, NY: John Wiley & Sons; 1950.
21. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley; 1968.
22. Guttman L. A basis for analyzing test-retest reliability. *Psychometrika*. 1945;10:255–282.
23. Crocker L, Algina J. *Introduction to Classical and Modern Test Theory*. Belmont, CA: Wadsworth; 1986.
24. DeVellis RF. *Scale Development: Theory and Applications*. 3rd ed. Los Angeles, CA: Sage; 2012.
25. Kline P. *The Handbook of Psychological Testing*. 2nd ed. London, UK: Routledge; 2000.
26. Messick S. Test validity and the ethics of assessment. *Am Psychol*. 1980;35(11):1012–27.
27. Freedman DR, Pisani R, Purves R. *Statistics*. 4th ed. New York: Norton; 2007.
28. Milewski GB, Sawtell EA. Relationships between PSAT/NMSQT scores and academic achievement in high school. New York, NY: The College Board; 2006. Report No.: R-N 2006-6.
29. The College Board. PSAT/NMSQT: what's on the test? https://collegereadiness.collegeboard.org/psat-nmsqt-psat-10/inside-the-test. Published 2011. Accessed June 22, 2018.
30. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York, NY: Houghton Mifflin; 2002.
31. Anastasi A. *Psychological Testing*. 6th ed. New York, NY: Macmillan; 1988.
32. Payne DA, McMorris RF. *Educational and Psychological Measurement: Contributions to Theory and Practice*. Waltham, MA: Blaisdell; 1967.
33. Himelfarb I. *Modeling the Change in the PSAT Scores: A Growth Modeling Approach*. Santa Barbara: University of California, Santa Barbara; 2012.
34. Yen W, Fitzpatrick AR. Item response theory. In: *Educational Measurement*. 4th ed. Westport, CT: Praeger; 2006:111–153.
35. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press; 1960.
36. Kolen MJ, Brennan RL. *Test Equating, Scaling, and Linking: Methods and Practices*. New York, NY: Springer; 2014.
37. Hambleton RK, Swaminathan H. *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer; 1996.
38. Masters GN. A Rasch model for partial credit scoring. *Psychometrika*. 1982;47:149–174.
39. Adams RJ, Wilson M, Wang W. The multidimensional random coefficients multinomial logit model. *Appl Psychol Meas*. 1997;21(1):1–23.
40. Akkermans W. Modeling sequentially scored item responses. *Br J Math Stat Psychol*. 2000;53:83–98.
41. Hoskens M, De Boeck P. A parametric model for local dependence among test items. *Psychol Methods*. 1997;2:261–277.
42. Wainer H, Bradlow ET, Wang X. *Testlet Response Theory and Its Applications*. Cambridge, UK: Cambridge University Press; 2007.
43. Thissen D, Orlando M. Item response theory for items scored in two categories. In: Thissen D, Wainer, eds. *Test Scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates; 2001:73–140.
44. Partchev I. Package "irtoys": a collection of functions related to item response theory (IRT). https://cran.r-project.org/web/packages/irtoys/irtoys.pdf. Published 2017. Accessed June 1, 2018.
45. Pukelsheim F. The three sigma rule. *Am Stat*. 1994;48(2):88–91.
46. Raykov T, Marcoulides G. *Introduction to Psychometric Theory*. New York, NY: Taylor & Francis; 2011.
47. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr Suppl*. 1969;17:1–100.
48. Gonzalez J, Wiberg M. *Applying Test Equating Using R*. New York, NY: Springer; 2017.
49. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: The Associations and Council; 2014.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-09-19 via free access