Research in English Language Pedagogy (2022)10(1): 1-28

©Author(s) 2022, open access at http://relp.khuisf.ac.ir/ DOI: 10.30486/RELP.2021.1945473.1321

Review article

Psychometrics Revisited: Recapitulation of the Major Trends in TESOL

Mohammad Ali Salmani Nodoushan

Institute for Humanities and Cultural Studies,Tehran, Iran

Submission date: 21 Nov, 2021 Acceptance date: 1 Dec, 2021

Abstract

A test is a tool for making quantified value judgments and/or comparisons, and a good test is a bias-free gauge that does its value judgments and quantifications with precision. This requires that the test be at least reliable. In applied linguistics in general, and in TESOL in specific, the question of test reliability has always been at the forefront of all test construction activities. As high-stakes gate-keeping tests gained more and more importance in a globalizing post-industrial world, the statistical procedures used to estimate their reliability indices, too, became more and more complex and precise. Classical Test Theory (CTT) is no longer preached, and test developers and testing agencies have resorted to Generalizability Theory (G-Theory) and Item Response Theory (IRT) as their main dishes; more recently, they have decided to spice up their activities with Differential Item Functioning (DIF). This paper seeks to provide the less-versed reader with a short and simple account of these topics. The aim of this paper is to turn the tumid prose describing complex mathematical and statistical topics in psychometrics and measurement into readable English so that students less versed in the field can make sense of them, and university professors can use the paper as a simple and informative source in their teaching activities.

Keywords: Classical Test Theory, Differential Item Functioning, Geralizability Theory, Measurement, Reliability, Testing

Corresponding Author's E- mail: m.nodoushan@ihcs.ac.ir

1. Introduction

To be able to compare objects/events to other objects/events for purposes of decision making, you will have to know, and be able to quantify, the attributes of those objects/events—albeit with precision (Salmani Nodoushan, 2020, 2021b). The term 'object' has been used in a loose sense here to include inanimate and animate entities, and it includes animals and human beings as well. The knowledge that enables you to quantify the attributes of an object/event can be called 'measurement', and the tools that you use for measuring those attributes are called gauges—tests, questionnaires, and scales included. That is, your gauges (e.g., a thermometer, a test, a questionnaire, etc.) can tell you which degree of the trait of interest the object/event being measured has achieved. The calculations that make your gauge—and whereby your quantifications—precise are called 'reliability estimates'.

Since its advent, measurement has remained a mainstay of all teaching and educational practices, and "obtained scores on the measurements plays [sic] a critical role in decision making about individuals and groups" (Holmes Finch et al., 2016, p. 1). Whenever decision-making is at stake, we need to make sure that our judgments are warranted and that the scores we rely on to make our judgments provide us with the best possible picture of our students' performance.

This is where psychometrics enters the game. Psychometrics is a subspecialty within educational psychology that has dovetailed (a) educational psychology and (b) statistics to ensure that any attempt at the development and vetting of measures (i.e., measurement tools, scales, tests, gauges, etc.) is done with precision and reliability. In other words, psychometrics brings an arsenal of statistical analyses (a) to bear on the precision of measures of performance and (b) to provide the researcher with detailed information concerning the reliability, precision, validity, and performance of gauges.

There are a good number of sources that describe psychometrics in detail, but it is cumbersome for most readers to read a huge number of sources to gain a modest knowledge of psychometrics. The huge amounts of redundancy and overlap that exist in and among these sources make it mandatory that a review article be written with the aim of cutting the long story of psychometrics short, and this is what the current paper has sought to achieve. This paper will review the main topics of psychometrics and describe their bearings on language testing and assessment.

2. Background

Measurement has been defined as the process of comparing an unknown quantity to a known and/or standard quantity (Pedhazur & Pedhazur Schmelkin, 1991; Salmani Nodoushan, 2009; 2021a). Any successful act of measurement requires the use of a well-constructed gauge put to its design-specific use/function in the right context. Gauges are scale-sensitive, and they cannot be put to use of one's own free will; one will have to follow certain procedures when one decides to implement them. In educational settings, tests are to be viewed as function-sensitive gauges that should not be used extravagantly; rather, they should only be used when all of the assumptions of their appropriate use have been met (Salmani Nodoushan, 2021a).

The term 'measurement' is frequently used in two different senses: (a) a general sense, and (b) a technical psychometric sense. In its general sense, measurement refers to the act of using a gauge to quantify a physical construct (e.g., length, height, weight, etc.) or an abstract trait (e.g., language proficiency, anxiety, motivation, etc.). In its technical psychometric sense, measurement refers to the idea of linking an observed value on a gauge to the unobserved construct/trait being measured. Perhaps it would be more precise if professionals had used two separate phrases to refer to these two senses of the term measurement: (a) internal measurement, and (b) external measurement. The former could be used to refer to all of the theoretical and statistical procedures that are followed in the process of developing a gauge, be it a test, a questionnaire, etc. The latter, by way of contrast, could be used to refer to the application of the already developed gauge for purposes of quantifying the trait for the measurement of which the gauge has been specifically devised. Perhaps the reason why this dichotomy has not been envisaged lies in the fact that internal measurement is a function of external measurement. In other words, a gauge (or scale) is first constructed based on certain theoretical assumptions, then it is put to pilot use to return some quantitative values (named quantitative data or scales), and then those values are statistically analyzed to show the internal properties of the gauge at hand (e.g., its validity, its reliability, its item difficulty and discrimination, etc.).

This means that when a gauge is put to the test, it returns some values. For example, when you use a speedometer to measure the speed of a car, it will show a number (i.e., a numerical value) that indicates how fast the car is moving. That number is an observed value—let's call it a datum, a quantification, or a scale. Note that the observed values on

any gauge are technically, and psychometrically, called the scales of the gauge—note that the term ‘scale’ is used in a general sense, too, to refer to a test, a questionnaire, or any other gauge that is used for measurement (Salmani Nodoushan, 2009). Psychometrically speaking, gauges (or ‘scales’ in the general sense) possess one of the four kinds of psychometric ‘scales’ (i.e., ‘scales’ in the psychometric sense): (a) nominal, (b) ordinal, (c) interval, and (d) ratio. These are differentiated from each other based on their properties (Bachman, 1990). Table 1 visualizes these scales and their properties.

Table 1.

Psychometric Scales and Their Properties

	Types of Scales			
Properties of Scales	Nominal	Ordinal	Interval	Ratio
Naming	+	+	+	+
Ordering	-	+	+	+
Equal distance	-	-	+	+
Absolute Zero Point	-	-	-	+

It was stated earlier that gauges are used to measure physical ‘constructs’ (e.g., speed, length, weight, heat, etc.) or psychological/behavioral ‘traits’ (e.g., competence, proficiency, self-esteem, tolerance for ambiguity, etc.). In acts of measurement in engineering and natural sciences, nominal properties of events and objects are not part of the measurement, but in soft sciences (e.g., education and psychology) and statistics, measures may include nominal scales as well (Pedhazur & Pedhazur Schmelkin, 1991). The academic discipline that pursues the production and dissemination of theoretical knowledge that has to do with the construction of gauges for the measurement of physical ‘constructs’ is called metrology. Psychometrics is the counterpart of metrology in soft sciences. As such, metrology is the science of measuring constructs, but psychometrics is the science of measuring traits. Seen from this perspective, traits are abstract constructs and constructs are physical traits.

All in all, psychometrics is the academic sub-discipline that pursues the production and dissemination of theoretical knowledge that pertains to the construction of gauges for the measurement of psychological/behavioral ‘traits’ (Salmani Nodoushan, 2009). Since its introduction to the field of language assessment, psychometrics has witnessed two major

paradigms: (a) Classical Test Theory (CTT), and (b) Item Response Theory (IRT). There is also a third camp—i.e., Generalizability Theory (GT or G-Theory)—which is mainly concerned with the validity of tests rather than their reliability (Bachman, 1990). I will return to these topics in the following sections.

3. Psychometrics in Language Assessment

In much the same way as a speedometer is a tool/gauge that measures the physical construct of 'speed', a test is a gauge that measures a psychological construct (i.e., a trait such as language proficiency, achievement, etc.). A trait is latent when it is concealed, abstract, and not open to direct perception. Examples of latent traits in language learning include proficiency, aptitude, etc. Since latent traits are not open to direct perception, they must be measured through the implementation of certain gauges. A language proficiency test, for instance, is a scale that is supposed to have been constructed in such a precise way as to show with maximum reliability and precision the size of the language knowledge that a test taker retains in his/her mind. In other words, the test is supposed to be a reliable measure of the trait it measures, and this suggests that it should be able to return the same observed value—with a tolerable amount of variation (or variances)—across repeated trials. Whenever the second, third or nth application of a measure returns a different observed value than its first application, the question that should be answered is if the observed variance has resulted from the improvement of the construct at hand (i.e., is a function of 'impact') or is due to measurement error (e.g., test takers' lapses in concentration, scorers' inconsistency, etc.). This is what reliability is all about—precision. When variations in observed values on a gauge across different trials are due to systemic changes (e.g., the improvement of the construct/trait the test measures), they are called systemic variances, but when they are due to random events (e.g., oversights, distractions, etc.), they are called unsystemic variances (Alderson et al., 1995)—please note that (un)systemic variances differ from (un)systematic variances to which I will return in my discussion of generalizability theory below. The more a test is able to measure systemic variance, the more reliable the test is. If a perfectly reliable test could be constructed, it would only measure systemic changes. This is possible in theory, but it has not been crystallized in practice yet.

Nevertheless, any act of measurement is always threatened by the presence of some degree of unsystemic (or systematic) variance. As such, reliability is an ongoing process,

and test constructors keep trying to find new ways of detecting and eliminating the causes of unsystemic—as well as systematic—variation. Uniform test administration, consistent marking, clarity of test rubrics and instructions, precise construct definition, and so forth are just a few strategies that boost reliability. All of the strategies and attempts aimed at boosting the reliability (and also validity) of measures have resulted in the emergence of two major measurement paradigms: Classical Test Theory (CTT), and Item Response Theory (IRT). Other camps have also appeared which include Generalizability Theory (or G-Theory) and Differential Item Functioning (DIF).

4. Classical Test Theory (CTT)

Imagine that you give your class a grammar test and rely on the scores the students gain on the test to tell how much grammar they know. If your test is a well-designed gauge, the test takers' scores should give you a trustworthy estimate of their knowledge of grammar. You also know that the scores are not free from error, and that there is always the possibility that some degree of error is present in any measurement. As such, the observed scores comprise both a true estimate of grammar knowledge and an estimate of measurement error. In other words, your observed variance comprises a true variance (also known as classical true score or CTS) and an error variance. You can therefore write the basic CTT equation $X = T + E$, where (a) X is the observed score (or the total variance) on the test for a given test taker, (b) T is his/her true score (i.e., systemic variance) on the trait being measured, and (c) E is random error (or unsystemic variance). This leads us to the basic assumption in CTT: Any observed score on a measure is a function of test takers' true and stable knowledge (i.e., systemic variance) and a set of ephemeral unstable and random factors (Haertel, 2006; Pedhazur & Pedhazur Schmelkin, 1991; Salmani Nodoushan, 2009; 2021a).

This suggests that it is assumed in CTT that T and E are uncorrelated. Feldt and Brennan (1989) argued that E in CTT belongs in one of the following four categories: (a) natural variation resulting from such factors as fatigue, hunger, mood, etc.; (b) environmental factors such as ambient noise, proctor's behavior, room temperature, etc.; (c) sporadic variations and inconsistencies in scores; or (d) malfunctioning items in the test booklet. No matter which of these factors is responsible for random E, its randomness implies several interesting points. First, if a test could be given to an individual repeatedly

and over a very large number of times, provided that (s)he forgot each time that (s)he had taken the test before, the mean of the errors across all administrations would be 0. In other words, the population mean would be $\mu_E = 0$. Second, the correlation between error variance *E* and true variance *T* would be 0. In other words, $r_{T,E} = 0$. Finally, error variances across multiple forms of the same test would also be uncorrelated. In other words, $r_{E1,E2} = 0$ (Holmes Finch et al., 2016).

All in all, CTT relies on three variables: *X*, *T*, and *E*. It also assumes that *X* is a composite variable in that it comprises *T* + *E*. Any time a variable is a composite of some other variables (e.g., *X* = *T* + *E*), its variance is a function of the sum of the variances of its components plus the multiplication of the covariance between its components by two. As such, the variance for *X* would be: $\sigma^2X = \sigma^2T + \sigma^2E + 2\text{cov}(T, E)$. Also note that standard deviation (*SD* or σ) is the square root of variance. Given the fact CTT assumes that *T* and *E* are uncorrelated, the covariance between them will be 0. As such, the composite variance of *X* can be rewritten as: $\sigma^2X = \sigma^2T + \sigma^2E$. This suggests that reliability is the ratio of the variance in *T* to the variance in *X*, and this can be represented in the following equation:

$$\rho_{xx} = \frac{\sigma^2T}{\sigma^2T + \sigma^2E}$$

Test reliability falls between 0 and 1, but no test can ever achieve complete reliability, so you would always expect a value smaller than 1 when you perform an estimation of test reliability (Harris, 1969). A test with a reliability index larger than 0.9 is excellent, but a range between 0.7 and 0.9 is also acceptable. CTT recommends that you compute the margin of error that you may want to accept in your measurements. This requires the computation of the standard error of measurement (*SEM*). Once you know the scores of a group of test takers, you can simply compute the group mean (*M*) and its standard deviation (*SD* or σ). To compute the mean, you can add up all scores (*X*s) and divide the result by the total number of test takers or population size (*N*). To compute the *SD* (or σ), you can use the following equation:

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

In this equation, x_i stands for each value from the population, Σ stands for sum total, μ stands for population mean, and *N* stands for population size. Also know that *SD* is the square root of variance (Harris, 1969). In other words, if you multiply *SD* by itself, the

result will be the test variance. Once you know the *SD* and the reliability of your test or measure, you can then use their values to compute the standard error of measurement (or *SEM*).

$$SEM = SD\sqrt{1-r}$$

The importance of *SEM* lies in the fact that it gives you the margin of error that you can expect in test takers' scores. As stated above, test reliability is never 100% complete, but you can compute the *SEM* and use it to make sure which score any test taker would gain on the test if (s)he took the same test many times (Harris, 1969). Psychometrically speaking, *SEM* shows the degree of confidence that a test taker's true score falls within a particular band score. If the total score on a multiple-choice test is 100 and the *SEM* for that test is 4, then you would say with 95% confidence that a test taker who has scored 60 on the test would score between 52 and 68 in 95% of his attempts if (s)he took the same test *n* times (e.g., 9500 times out of 10000 attempts). As such, band scores in CTT are ±2*SEM*s from observed scores (Harris, 1969). The 95% confidence comes from 'the ±2*SD*s from the *M* formula' based on the area under the normal probability curve. Figure 1 visualizes the concept of normal distribution.

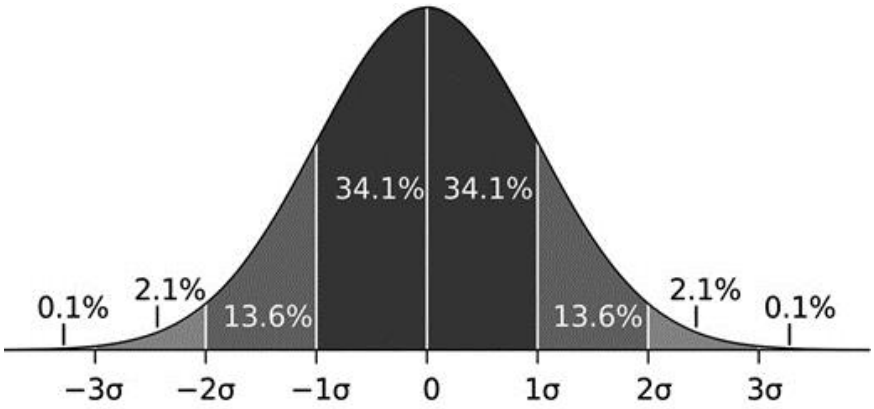

Figure 1. Areas Under the Curve in Normal Distribution (Adopted from Harris, 1969)

Nevertheless, reliability is just one of the qualities of a good test. CTT holds that the most important quality of a test is its usefulness (i.e., the use for which it is intended or its function). Test usefulness is a function of test reliability, test validity, and test practicality (Bachman, 1990). Together, they have been referred to as the '*sine qua non*' of a test (Salmani Nodoushan, 2020). To guarantee the usefulness of a test, CTT resorts to several important concepts including among other things (a) the Guttman's lower bounds to

reliability, (b) the reliability index, (c) the correction for attenuation, (d) the Kuder-Richardson formulas, and (e) the Spearman-Brown formula (Salmani Nodoushan, 2009).

Anyway, the forces that went hand in hand in the early 20th century to bring about CTT were threefold: (a) acceptance of the presence of error in any measurement also known as random latent variable, (b) acceptance of its random nature, and (c) a conception of correlation and how to index it (Salmani Nodoushan, 2009). Some would argue that CTT began in 1910 when Spearman (1910) proposed the concept of correction for attenuation (Bachman, 1990). As already noted, CTT was connected to two concepts in the previous paragraphs: (a) reliability, and (b) *SEM*. In fact, these two concepts are the main offshoots of CTT (Salmani Nodoushan, 2009).

CTT has its original roots in the ideas of Galileo, the Italian scientists of the 17th century, who believed that errors in scientific observations (a) were inevitable, (b) were distributed symmetrically, and (c) tended to cluster around their true values (Salmani Nodoushan, 2009). It was not until the beginning of the 19th century when scientists mainly astronomers—began to emphasize the importance of studying errors. As a result of the upsurge of interest in the study of errors, Carl Friedrich Gauss proposed the concept of normal distribution (Figure 1), and around the turn of the 20th century, measurement specialists had already accepted the idea of measurement error. It was around this time when measurement specialists, and specifically Charles Spearman, proposed the notion of correction of a correlation for attenuation—also known as correlation disattenuation or the disattenuation of correlation (Osborne, 2003). Spearman, for example, realized that the absolute value of the correlation coefficient between the different measurements "for any pair of variables must be smaller when the measurements for either or both variables are influenced by accidental variation than it would otherwise be" (Salmani Nodoushan, 2009, p. 2).

Out of these considerations came the idea of test/measure reliability. Reliability is defined in terms of the consistency and precision of measurement. A reliable test score is one that remains consistent across different characteristics of the testing situation (Bachman, 1990; Brown & Salmani Nodoushan, 2015). Validity, another quality of a good test, has to do with the idea of test function. Some would argue that validity is obtained when reliability is squared, but the picture is not that simple; not all instances of common variance can be taken as estimations of validity (Salmani Nodoushan, 2009). To be valid, a test is supposed to measure the trait it has been constructed to measure. Otherwise, it

would be invalid. Nevertheless, reliability and validity work in tandem in that a test cannot be valid unless it is first and foremost reliable (Bachman, 1990; Harris, 1969). It is noteworthy that all estimations of reliability are done with the aim of maximizing true variance (i.e., classical true score or CTS) and minimizing error variance. As Harris (1969) and Bachman (1990), among many others, have ardently argued, investigations of reliability might be based on (a) logical analyses done with the aim of identifying error sources, or (b) empirical studies aiming at estimating the magnitude of the impact of errors on test performance. CTS is essentially the offspring of these assumptions (Salmani Nodoushan, 2009). In addition to envisaging two uncorrelated sources of variance (i.e., true variance and random error variance), CTS also holds that true score variations are a function of disparities in testees' ability levels (Salmani Nodoushan, 2009).

As for reliability, CTS has envisaged three reliability models: (a) estimates of internal consistency or item-total correlations, (b) estimates of through-time stability, and (c) estimates of equivalence among the parallel forms of the same measure. Estimates of internal consistency reveal if test takers have performed consistently on different parts of a measure or test. Estimates of through-time stability determine if a test taker's *T* remains stable over repeated administrations of the same test. Finally, estimates of equivalence among the parallel forms of the same measure reveal if different forms of a test that are highly correlated return the same CTS for any given test taker (Harris, 1969; Salmani Nodoushan, 2009). The basic assumption behind the development of highly correlated parallel forms of a measure is that the error as well as the true variances of one form will equal those of any the other form. Metaphorically speaking, this is similar to the idea that different speedometers manufactured by different factories will inevitably show the same value for the speed of a car once they are installed in that car.

The three reliability models just described (i.e., equivalence, through-time stability, and internal consistency) have turned up into several procedures for the estimation of reliability in CTT. As for the 'estimates of internal consistency' model, seven types of reliability estimation have been documented in the existing literature: (a) split-half reliability, (b) Spearman-Brown split-half estimate, (c) the Guttman split-half estimate, (d) Kuder-Richardson reliability coefficients, (e) coefficient *alpha*, (f) intra-rater reliability also known as 'regrounding', and (g) inter-rater reliability (cf., Salmani Nodoushan, 2009). The last two procedures may also be collectively called 'rater consistency'. It should be

noted that in split-half reliability, a single test is broken into two halves through (a) the odd-even, (b) the first-half-second-half, or (3) the random-halves method—with the assumption that the two halves are both locally independent and totally equivalent—and then the correlations between the two halves is computed (Harris, 1969).

Nevertheless, splitting a measure into two halves reduces the total test length, therefore the test specialist must make up for this (a) either through the Spearman-Brown prophecy formula, which assumes the two halves to be equivalent and also experimentally independent, or (b) through Guttman split-half estimate, which does not make such an assumption (Salmani Nodoushan, 2009). Since it does not assume that the two halves are equivalent and experimentally independent, the Guttman's formula may be used to estimate the reliability of the whole measure directly. It should be noted that the application of the Spearman-Brown prophecy formula to two unequal halves turns up into an underestimation of reliability, and also that its application to experimentally dependent halves returns an overestimation of reliability. Perhaps a better strategy for the scorer would be to take these consecutive steps: (a) engage all of the splitting methods, (b) estimate the reliability coefficient for any of them, and (c) find the average of these coefficients (Salmani Nodoushan, 2009). Nevertheless, it is often advised that scorers use the Kuder-Richardson formulas to avoid the potential errors that lie within the split-half method.

Split-half reliability may be a suitable method for multiple choice tests where there are enough items to assign to two halves, but there are also tests that do not lend themselves readily to split-half reliability estimation—often because their different sections are not locally independent. In such cases, 'test-retest reliability' might be a better option, but it is not free from complications in that a short interval between the two administrations might expose the measure to practice effect—also known as carry-over or history effect—and a long interval might expose it to the effect of ability change (Salmani Nodoushan, 2009); hence, the test-retest reliability dilemma. In fact, this dilemma has motivated the use of equivalent or parallel test reliability estimates. All in all, CTS reliability estimates are always prone to error because (a) there may be some kind of interaction among the different sources of error, (b) some error sources might not be controllable, (c) the estimation of error sources might be relative, (d) errors are treated in a homogeneous way while they might not be genuinely homogeneous, and (e) errors are

taken to be random, but not systematic (Salmani Nodoushan, 2009).

Unlike reliability which has to do with the size of variation in scores resulting from (a) test method facets and (b) systematic and random measurement errors, validity has to do with the relationship between test performance and performance in non-test contexts. CTT approaches test validity from three perspectives: (a) content validity, (b) criterion related validity, and (c) construct validity. Content validity evaluates the correspondence between the content of a test and the content of the corpus the test claims to measure; in other words, it correlates test specifications and test content, guarantees measure accuracy, and prunes out any potentially harmful backwash (Salmani Nodoushan, 2021b). Criterion related validity, on the other hand, correlates two different tests that claim to measure the same trait (e.g., the *TOEFL* and the *IELTS*). Criterion related validity has two manifestations in CTT: (a) concurrent validity, and (b) predictive validity. When the test and the criterion to be correlated are administered at about the same time, concurrent validity is at stake, but in predictive validity, which concerns the degree to which a test can predicate candidates' future performance, the test and the criterion are subsequent to each other (Valette, 1977). Finally, construct validity inspects the underlying structure of a test to see if it measures the predefined ability it claims to measure (Harris, 1969); the test, its parts, and the testing technique are said to possess construct validity only when the test gauges what it has been specifically devised to gauge. It should be noted that the term 'construct' refers to the latent trait of interest (e.g., language aptitude, language proficiency, etc.) that the test claims to measure.

In addition to content validity, criterion related validity, and construct validity, CTT also talks about 'face' validity, which concerns the appearance of a test. A test has face validity if it appears to be testing what it claims to be testing. Although face validity is hardly a strictly scientific concept, it is very important. Any test that lacks face validity is not taken seriously by test users (i.e., testees, instructors, educators, and so forth). As such, the notion of face validity tacitly implies that new testing techniques, especially indirect ones, may fail to convince test users.

5. Generalizability Theory (G-Theory)

As stated above, CTT is based on a true variance (or CTS) and a random error. It also relies heavily on coefficients of correlation. Nevertheless, later developments in

psychometrics showed that CTS might itself be overestimated because of the presence of what has come to be known as non-random or systematic error. As such, a new perspective on evaluating reliability which is linked more intimately to the fundamental equation in CTT (i.e., $X = T + E$) has been proposed and has come to be known as generalizability theory (GT). It was stated above that in CTT, reliability is defined in terms of the ratio of true score variance (σ^2T) to observed score variance (σ^2X). CTT assumed a linear and reverse relationship between reliability and error variance, "such that greater measurement error would be associated with lower reliability" (Holmes Finch et al., 2016, p. 76).

In spite of its being totally based on the conception of measurement error, CTT did not make any modest and tangible attempt to quantify measurement error. By way of contrast, GT addresses measurement error directly and seeks to estimate its magnitude. It then dovetails (a) the estimated error magnitude and (b) the information about the observed score to compute its own *g* coefficient—which is an estimation of reliability (Holmes Finch et al., 2016). GT does not see *X* as a function of only *T* and random *E*, but argues that *T* itself may be contaminated in that it may also comprise some form of systematic error variance which is hidden due to its non-random and systematic nature. As such, *T* is inspected and the systematic error hidden in it is found and eliminated (Holmes Finch et al., 2016; Shavelson & Webb, 1981; Shavelson et al., 1989). GT considers the universe score to be a function of (a) systemic *T*, (b) systematic *E*, and (c) residue or random *E* (Cronbach et al., 1972). As such, "a given measure or score is a sample from a hypothetical universe of possible measures," and "a score is a multi-factorial concept" (Salmani Nodoushan, 2009, p. 6).

GT, just like CTT, has its own set of professional nomenclatures developed for purposes of the precise description of its practices. GT assumes that an observed score obtained from the administration of a measure comes from a 'universe' of possible scores for the same testee on a given test. As such, a test used to measure a trait is just one of the many possible instruments that could have been used to measure the trait. All of the imaginable variables that might play a role in the measurement are called 'facets' (or objects of measurement); the number of items, the test rubrics, the temperature of the testing room, the test takers' moods and other physical and mental characteristics, the proctors' physical and behavioral characteristics, topastic or guessing error, raters (e.g., judges, teachers), psychological task set, measurement occasion (or testing time), the

instrument type (e.g., portfolios, writing samples, multiple-choice tests), cultural content, and so forth are all 'facets', and a 'universe' score is a function of multiple facets (Holmes Finch et al., 2016). The basic assumption in GT is that each of the many facets that can be pinpointed has a unique share of the observed variance, and GT aspires to estimate the exact share of each and every facet (Shavelson & Webb, 1981). GT does this—i.e., the estimation of the relative contribution of each facet to the total score—in what has been called a 'G-study' (or generalizability study). In other words, a G-study isolates the unique share of each and every facet through a 'variance components analysis' (i.e., through decomposing the observed score into its constituents) and explains its contribution to the observed score (i.e., total variance) in the form of explained variance (Holmes Finch et al., 2016). Once facet-specific variances are calculated, they are used in a 'D-study' (or a decision study) which aspires to generalize the obtained results to the 'universe' of interest—also known as the 'universe of generalization'. A D-study also makes it possible for the researcher to estimate the "relative reliability of the measure under different conditions with respect to the facets" (Holmes Finch et al., 2016, p. 77).

Perhaps Ebel (1951) had the greatest impact on the development of GT. In his article on the reliability of ratings, Ebel (1951) identified two sources of errors: (a) rater main effects included, and (b) rater main effects excluded. He grappled with this issue until GT was fully formulated, and only then could he distinguish between 'relative' and 'absolute' errors in various factorial designs (cf., Kane & Brennan, 1980). Later, Lord (1957) suggested a 'binomial error model' which has since been an integral part of GT. In lay terms, GT has upcycled the notion of *SEM* from CTT to estimate 'conditional' standard errors of measurement (*CSEM*s).

It was stated earlier that in GT "a given measure or score is a sample from a hypothetical universe of possible measures," and "a score is a multi-factorial concept" (Salmani Nodoushan, 2009, p. 6). This suggests that generalizations from a single measure can be made to a universe of measures. This implies that reliability is in essence a matter of Generalizability. Nevertheless, making generalizations requires that we define our universe of measures with precision. All in all, GT (a) specifies all of the objects of measurement, (b) measures their magnitudes, and (c) generalizes from there to the universe of generalization. In other words, GT brings one measure from a universe of measures to make its own calculations, and then implements those calculations to test interpretation and use.

Needless to say, this is a 'validation' issue. Each facet within the universe of generalization has (a) its own unique characteristics and (b) varying conditions (Salmani Nodoushan, 2009). These, in turn, contribute to the overall variance obtained from a measure such that the generalizability coefficient computed is dependent on them. As such, the generalizability coefficient is "the proportion of observed score variance that is universe score variance" (Salmani Nodoushan, 2009, p. 6). It was stated earlier that the objects of measurement in GT include (a) universe score variance, (b) systematic variance, and (c) random or residual variance. Since GT identifies two sources of error (i.e., systematic and random), it is not surprising that a generalizability coefficient is often smaller than its CTT counterpart (i.e., reliability coefficient). Nevertheless, a generalizability coefficient is always more precise than a reliability coefficient in that the former is the end product of a process that has located systematic variance and eliminated it (Brennan, 1984; Cronbach, 1951, 1984; Fisher, 1925; Lindquist, 1953; Salmani Nodoushan, 2009). It should also be noted that GT adopts different approaches to reliability in criterion referenced (or domain referenced) versus norm-referenced measurements.

Very often it is claimed that GT has blurred the distinction between validity and reliability. However, the reality is that only a small portion of the GT literature directly relates to validation. Perhaps the first attempt at linking GT to validity was made by Kane (1982), who proposed the 'sampling model for validity' as a seminal contribution to GT. Kane's model links GT to key issues that were subsumed under validity in CTT (Salmani Nodoushan, 2009). Unlike CTT, GT does not view validity as a three-dimensional concept; rather, it takes validity to be a unitary concept at the heart of which lies the notion of construct validation (cf., Messick, 1988; Salmani Nodoushan, 2020). Construct validation viewed through the GT lens engulfs (a) convergent and (b) discriminate types of evidence that work in tandem. This unitary concept of validity comprises the following components: (a) the content aspect, (b) the substantive aspect, (c) the structural aspect, (d) the generalizability aspect, (e) the external aspect, and (f) the consequential aspect (Salmani Nodoushan, 2009). Table 2 illustrates these validity components.

Table 2.

Facets of Validity Envisaged by Samuel Messick (1988)

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/utility
Consequential Basis	Value Implications	Social Consequences

As indicated by Table 2, Messick's (1988) four-way framework of validity stands on a two-fold pedestal: (a) an evidential basis, and (b) a consequential basis. Together, these bases justify how a measure should be used and/or interpreted. This framework is based on the assumption that (a) relevance, (b) utility, and (c) construct validity should work in unison, and on an evidential basis, in the process of test use (Salmani Nodoushan, 2009). Likewise, construct validation and value implications are part and parcel of the process of test interpretation. Similarly, the consequential perspective on test use integrates (a) test relevance, (b) its construct validity, (c) its utility, and (d) its social consequences. All in all, such a unitary approach to test validation holds that validity is a function of evidential and consequential bases, and that these bases, in turn, are sensitive to (a) content relevance, (b) criterion relatedness, and (c) construct meaningfulness (Salmani Nodoushan, 2009). Validity, seen from this perspective, is an integrated evaluative judgment about the degree to which "empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *attitudes* based on test scores or other modes of assessment" (Messick, 1988, p. 13); it relates to the available evidence that can support test interpretation and the potential consequences of its use (Salmani Nodoushan, 2009).

All in all, GT capacitates researchers to cope with measurement error directly, and to compute its magnitude. They can then bring their estimations to bear on estimations of reliability. A clear advantage of GT over other estimations of reliability—such as *alpha*—is its capacity to isolate and quantify multiple sources of error.

6. Item Response Theory (IRT)

Perhaps the main difference between Item Response Theory (or aka IRT) and CTT is that the focus of the former is at the item level whereas the focus of the latter is at the whole test (or scale) level (van der Linden & Hambleton, 1997; Yen & Fitzpatrick, 2006). The suite of IRT tools contains a set of statistical models that aim to detect measurement error at item level. In essence, IRT focuses on the relationship among (a) the testee, (b) the

test items, and (c) the probability of the testee providing a given response (e.g., incorrect or correct) to the item (Holmes Finch et al., 2016). Models of IRT have been developed to tackle (a) dichotomous and (b) polytomous items. Dichotomous items have two possible answers (correct or incorrect shown in 0s and 1s). Polytomous items have more than two responses (e.g., Likert type items). IRT models are based in the logistic framework and are differentiated in terms of the amount of information they contain about test items (Holmes Finch et al., 2016).

Item Response Theory is also known as 'aka IRT' or latent trait theory. The true score in IRT is defined on the latent trait at hand rather than on the test (Salmani Nodoushan, 2009). IRT is quite often engaged in different activities including (a) item bias analysis, (b) equating, and (c) tailored testing—among other applications. Item bias analysis reveals if a test item is functioning without bias (i.e., DIF or systematic error) across testee groups (e.g., males versus females, blacks versus whites, etc.). Equating is very much similar to what GT aspires to do—generalizing from an observed score to the universe; in other words, the score on one test can be used as the basis for predicting the equivalent score on another test—which is roughly similar to the CTT notion of expectancy tables. Finally, tailored testing dispenses with the idea of giving different testees the same test for purposes of ranking them; no matter which combinations of different test items are given to different testees, IRT capacitates the test maker to place test takers on the same scale (Salmani Nodoushan, 2009). This can boost test security because each individual can receive a different set of items, but still be comparable to other individuals (Bachman, 1990).

The basic one-parameter IRT model—also known as the 1-parameter logistic (1PL) model—is based on the works of Rasch (1980)—note that Rasch is just a special case of 1PL—and assumes that a testee's performance on a test item is a function of the difficulty of the item and the testee's ability level. By way of contrast, many-facet IRT models allow parameters other than item difficulty and testee's ability level (e.g., rater severity) to be included as assessment variables in estimating the testee's underlying ability (Salmani Nodoushan, 2009). IRT models are essentially tailored to the estimation of test reliability, but they are also sometimes used for the estimation of validity. This latter use has been criticized by some measurement experts who believe that IRT assumptions (e.g., unidimensionality) do not allow IRT models to be engaged in validity estimations (cf., Alderson et al., 1995).

IRT is based on probability theory. As such, in cases where the difficulty level of an item is the same as a testee's ability level, the testee will have a 50/50 chance of getting that item right. Theoretically speaking, this allows students' scores and item totals to be transformed on to one scale in such a way as to make them related to each other (Salmani Nodoushan, 2009). This theoretical relationship between testees' actual item performance and the abilities that underlie item performance is visually described in what has come to be known as an Item Characteristics Curve (ICC)—also known as an item trace. It is a common tool implemented to examine the properties of individual test items. An ICC "relates the latent trait being measured (on the *X* axis), with the probability of a correct response (in the case of dichotomous items) based on the particular model selected on the *Y* axis" (Holmes Finch et al., 2016, p. 6).

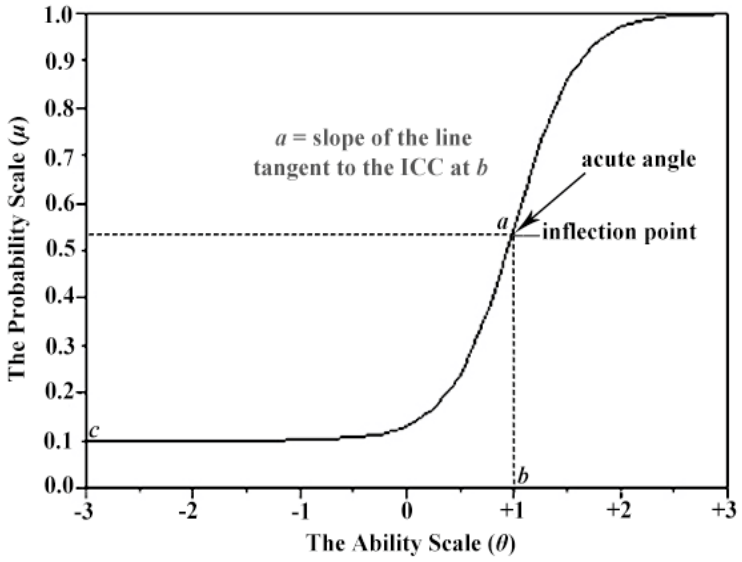

Figure 2. The Appearance of a Typical ICC for a Hypothetic Test Item

In much the same was as a physician can develop an idea of how a patient's heart is functioning just by looking at an electrocardiogram (ECG), an IRT expert can look at an item trace to develop an idea of how a testee has performed on a test item (van der Linden & Hambleton, 1997; Yen & Fitzpatrick, 2006). In the case of an ECG, the functioning of the patient's heart is a function of his or her level of health. Likewise, in the case of an ICC, the testee's item performance is a function of his or her level of logit scale ability symbolized as theta (θ)—also known as latent trait (Yen & Fitzpatrick, 2006). It has been

agreed by convention that the θ scale should be set by assuming a mean of zero and a *SD* of one for the population—which is analogous to the normal probability curve values or traditional z-scores in CTT (Salmani Nodoushan, 2009). As such, the values on the ability scale in the ICC in Figure 2 are tantamount to *SD*s from the mean of 0. The probability scale in the ICC indicates the probability of getting the item right. Figure 2 displays the overall appearance of a typical ICC for a hypothetic test item.

ICCs are the stepping stones of all IRT models. ICCs yield three basic parameters for test items: (a) the discrimination parameter or item discrimination symbolized by *a*, (b) the item difficulty/facility parameter symbolized by *b*, and (c) the topastic or guessing effect symbolized by *c*. IRT also takes the *X* axis to be the ability scale (symbolized by θ, and the *Y* axis to be the probability scale symbolized by μ or *p*. The slope of the curve tangent at *b* is the *a* parameter. This can be computed very easily. If you draw a line parallel to the *X* axis at the inflection point and another line flush with the curve at the inflection point, their intercept will form an acute angle the tangent of which you can compute using simple math. There is also a constant value equal to 02.71 and symbolized by *e*. Once the values for these parameters are known, they can be fed into the following equation:

$$\mu = c + (1 - c)\frac{1}{1 + e^{-1.7a(\theta - b)}}$$

The most important parameter in IRT is the *b* parameter which shows item difficulty. It is this parameter that sets the location of the inflection point of the ICC. If you drop a vertical line from the inflection point to the *X* axis, your *b* value will be found. The *c* parameter is the tangent of the point at which the ICC intercepts the *Y* axis. In ICCs where the *c* parameter is equal to zero (i.e., there is no topastic effect), the inflection point on the curve will be at μ = 0.50 (Salmani Nodoushan, 2009). The one-parameter IRT model (also known as the 1PL model) is sensitive to only the *b* parameter and only lets this parameter vary.

As stated above, the steepness of the ICC at its steepest point marks the *a* parameter, which shows item discrimination defined in terms of the relationship between items and individuals' ability levels. Remember that the *a* parameter is found by taking the slope of the line tangent to the ICC at b (Salmani Nodoushan, 2009). The closest relative of the *a* parameter in CTT is item total correlation. The steepness of the curve has a direct relationship to item discrimination such that the steeper the curve, the more discriminating the item and the larger the *a* parameter. In other words, the more the *a* parameter

decreases, the flatter the curve gets and the less discriminating the item will be (Salmani Nodoushan, 2009). Just like CTT where items with an item discrimination smaller than 0.3 or larger than 0.9 were discarded from the test (Harris, 1969), items with very low or very high *a* parameter values in IRT are not useful. Perhaps it would not be wrong to claim that IRT is essentially an item-level approach to the computation of item discrimination indices whereas CTT used a scale-level approach to the computation of those indices. Remember from Harris (1969) that, in CTT, you would (a) rank the total scores of your testees from high to low, (b) separate 27% from each end and label them *H* and *L*, and (c) use the following equation to compute item discrimination indices:

$$ID = \frac{No.of\ correct\ H - No.of\ correct\ L}{N\ of\ either\ H\ or\ L}$$

Unlike the one-parameter IRT model which is (a) sensitive to only the *b* parameter and (b) only lets the *b* parameter vary, the two-parameter IRT model allows both *a* and *b* parameters to vary; that is, the 2PL model engages both *a* and *b* in its descriptions of scale items.

In three-parameter models, a third parameter enters the computations. It is the *c* parameter, which is essentially an index for guessing or topastic effect—also known as pseudo-chance; hence, the guessing parameter. Using the professional terminology of projective geometry, we can say that the *c* parameter is a lower asymptote—i.e., a line which is tangent to the ICC at a point at infinity (Nunemacher, 1999). In lay terms, the *c* parameter is "the low point of the curve as it moves to negative infinity on the horizontal axis" (Salmani Nodoushan, 2009, p. 10). The *c* parameter quantifies the possibility for a low-ability examinee (say a chicken) to provide the correct answer to an item. As such, it might be used to model topastic effect in multiple-choice test items.

All in all, each and every item has its own item-specific ICC in IRT. Depending on which IRT model you use to plot your ICCs, you will gain access to different types of information about your test items. If you use the 1PL model, you will only gain information about item facility. Note that the 1PL model assumes that all test items have equal values for item discrimination—which equals 1 in the case of the Rasch model, which is a special case of the 1PL, à la Embretson and Reise (2000) and de Ayala (2009); this assumption is what differentiates 1PL from 2PL and 3PL models (Holmes Finch et al., 2016). If you use the two-parameter model, your ICCs will afford information about both item facility and item discrimination. Finally, ICCs from a three-parameter model will tell

you a lot about item facility, item discrimination, and guessing. ICCs are plotted on the basis of two scales: (a) the *X* axis or ability scale, and (b) the *Y* axis or the probability scale (Hambleton & Swaminathan, 1985). In sum, you can use the best-fitting IRT model for two purposes: (a) to assign estimates to test items, and (b) to assign scores to test takers. The scores you assign to test takers through the implementation of the best-fitting IRT model can be called 'ability' scores.

When items are polytomous, they do not have a binary response (i.e., incorrect/correct or 0/1). Unlike dichotomous items, polytomous test items are not scored dichotomously; rather, their responses are gradable on a cline (e.g., on a Likert scale). In other words, polytomous items with graded responses may include several categories (e.g., poor, fair, good, and excellent). In such cases, the IRT models described above cannot be used to model item responses (Holmes Finch et al., 2016). As such, new more complex models have been developed in measurement and psychology that can handle polytomous test items (e.g., the generalized partial credit model, or GPCM) a discussion of which is beyond the scope of the current paper. For more on polytomous IRT models, please see Muraki (1992).

It has been claimed in the literature on IRT that ability scores are precise measures of examinee's real abilities (Hambleton & Swaminathan, 1985). In other words, IRT claims to guarantee the precision of measurement. Nevertheless, CTT and GT have both been criticized on the assumption that their estimations of reliability, generalizability, and *SEM* are not precise due to the fact that these estimations are sample-dependent. When tests are sample-dependent, they find different reliabilities when different sample groups take them. Another criticism leveled against both CTT and GT is their treatment of error variance as homogeneous across individuals. This approach to error variance is in itself a source of measurement imprecision. IRT claims to have eliminated both of these sources of imprecision (Salmani Nodoushan, 2009).

Anyway, ICCs in IRT yield different types of information. Each item is therefore said to have an information function which can technically be called Item Information Function (IIF), which is based on two pillars: (a) ICC slope, and (b) variation at each ability level. An IFF comprises the amount of information the item affords for the estimation of any test taker's level of ability. Once you have the IFFs for all items in a test, you can sum up all of them to arrive at the Test Information Function (TIF), which is an

estimate of how much information your test yields at different ability levels. As such, the *SEM* for each ability level will be the reverse of TIF for that ability level. This quality of TIF in IRT (a) makes IRT measures sample-independent, and (b) results in measurement precision and reliable reliability (Salmani Nodoushan, 2009).

Nevertheless, IRT models have also their downsides, and they have indeed been criticized on the ground that they are not that much applicable when it comes to the estimation of validity indices (Salmani Nodoushan, 2009). Furthermore, the basic assumption of IRT models—the fact that they envisage a single latent trait (or their unidimensional assumption)—is also a bone of contention. As such, the unsatisfactory theoretical assumptions of IRT models have made their application to the estimation of validity invalid. As stated above, the qualities of a good test include validity as well as reliability; if IRT models are invalid for the estimation of validity, IRT cannot be considered as a comprehensive and exhaustive theory of measurement. As such, measurement specialists will need to work on this aspect of IRT.

7. Differential Item Functioning (DIF)

It was stated above that IRT (a) might be used for item bias analysis and (b) has been criticized for its lack of attention to validity. Models of differential item functioning (DIF) have come about to make up for this (Karami, 2018; Karami & Salmani Nodoushan, 2011). DIF holds (a) that test score use across different populations (e.g., students, patients) and diverse settings (e.g., educational, vocational) has different implications, and (b) that this makes the correct measurement of the intended trait vital (Holmes Finch et al., 2016). In other words, test scores must be valid for the target use for which they were designed in the first place (Linn, 2009). As such, DIF analysis—which is a more recent development of IRT—addresses the question of whether test items are fair and appropriate (i.e., valid) for assessing the traits of all test takers regardless of their subgroup membership. In essence, DIF analysis aims (a) at eliminating potential unfairness—i.e., potential presence of DIF in test items—in score-based decision making (Wu et al., 2007; Zumbo, 1999) and (b) at ensuring the "equivalent meaning of test scores across diverse groups" (Holmes Finch et al., 2016, p. 196).

The basic assumption in IRT is that all test takers who have been matched on the latent trait of interest should have the same probability of getting the test items right, but

where DIF is present in test items, this basic assumption is violated (Camilli & Shepard, 1994). In the literature on DIF, it has been claimed that group specifications (e.g., gender, language spoken in the home, race/ethnicity, opportunity to learn, other testee demographics, or their intercepts) may be responsible for bias in test items, so DIF analysis is vital for the detection and elimination of item bias.

Where DIF is present in an item, it may be either (a) uniform or (b) nonuniform. The former has to do with cases in which "the probability of a correct item response between two matched groups differs *consistently* across the entire range of ability" or the ability axis (Holmes Finch et al., 2016, p. 196, italics mine). By way of contrast, nonuniform DIF is present in cases where "the probability of a correct item response differs according to the groups' standing on the latent trait" (Holmes Finch et al., 2016, p. 197). Figure 3 is a visual representation of uniform and nonuniform DIF.

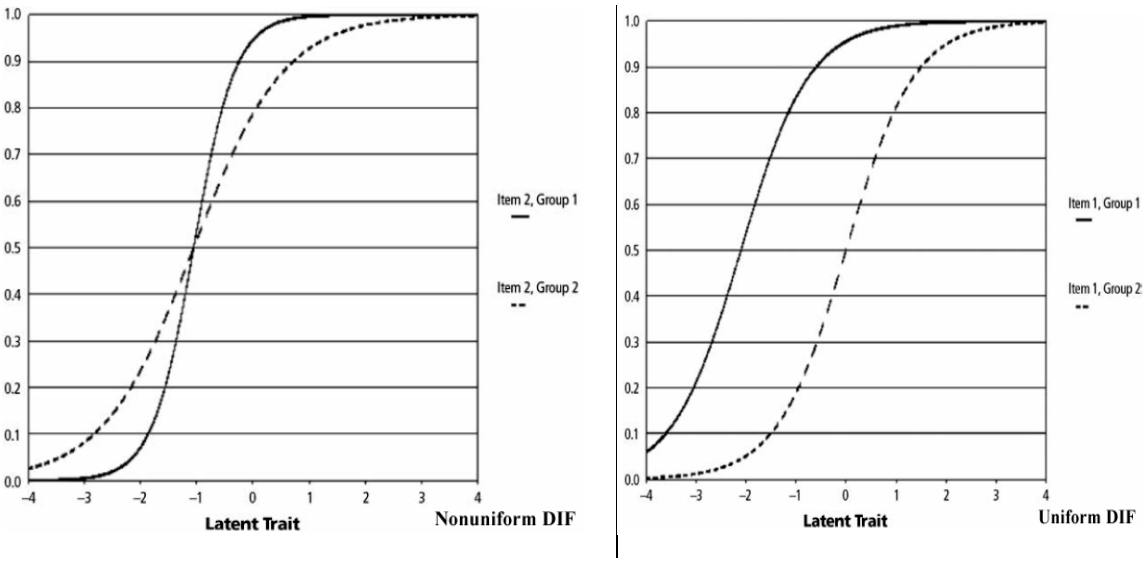

Figure 3. Visual Representation of Uniform and Nonuniform DIF (adopted from Holmes Finch et al., 2016, p. 197, pp. 197, 198)

DIF is inevitable when different groups of test takers with the same latent trait (the same level of ability/skill) have diverse probabilities of giving a certain response to an item. As such, DIF is technically described as differences in item response probabilities (Salmani Nodoushan, 2009). Raju (1988) has argued that DIF is a difference in the parameters of the IRT model between the groups being tested (cf., Holmes Finch et al., 2016). Since different IRT models for dichotomous items discussed above include different

item properties (i.e., *a*, *b*, and *c* parameters), Raju's argument suggests that, depending on the IRT model chosen for plotting ICCs, dichotomous items may show *a*-DIF, *b*-DIF, *c*-DIF, or a combination of these (cf., Finch & French, 2011).

A point of caution is that 'impact' and 'DIF' are distinct and should not be mistaken. Impact, just like DIF, is a difference in item performance between groups of testees, but this difference, unlike in DIF, is a function of a 'desired' disparity in the size of the latent trait that the item measures (Clauser & Mazor, 1998). In true experimental designs, for instance, we expect a disparity between the experimental and control groups, and we call it impact. DIF, on the other hand, is undesirable in that it is a function of some intervening variable that we have not been able to eliminate from our measurements. In lay terms, impact is roughly similar to 'systemic' variance in GT, but DIF is roughly similar to 'systematic' variance in GT. As such, item impact and DIF can be differentiated in terms of the matching of individual test takers on the latent trait of interest (Holmes Finch et al., 2016). In other words, before any DIF analysis, you must match your test takers on the latent trait, and this is by no means a trivial matter.

Once you have matched the test takers on the latent trait, you can use certain DIF analysis procedures to detect the presence of DIF in your test items. The most popular method for small populations (and mainly for dichotomous items) is the Mantel–Haenszel chi-square test (Camilli & Sheperd, 1994; Holland & Thayer, 1988; Mantel & Haenszel, 1959). For larger populations (and mainly for polytomous items), the Logistic Regression method is preferred to the Mantel–Haenszel chi-square test. A discussion of these methods is beyond the scope of this paper, but the interested reader is invited to see Agresti (2002), Clauser et al. (1993), Cohen (1992), Donoghue and Allen (1993), Jodoin and Gierl (2001), Karami (2018), Karami and Salmani Nodoushan (2011), Michaelides (2008), Narayanan and Swaminathan (1996), Rogers and Swaminathan (1993), Roussos and Stout (1996), Swaminathan and Rogers (1990), Thomas and Zumbo (1996), and Zwick (2012). All in all, DIF analysis is in essence on a par with GT in that it, just like GT, seeks to detect sources of systematic variance that are latent-trait irrelevant.

8. Conclusion

The field of language testing, as we saw in this paper, has been informed by developments in educational psychology which, in turn, has been informed by statistics,

projective geometry, and mathematics. The marriage between educational psychology and statistics has resulted in the emergence of psychometrics, which has flourished since its advent in 1910 to embrace a rich repertoire of professional terms and analytical methods. Reliability and validity are the main goals of psychometric analysis, and four different perspectives have to date been adopted in psychometrics to perform reliability and validity analyses: CTT, IRT, DIF, and GT.

Nevertheless, what was presented in this paper is not the sole property of language assessment and testing. No matter where a testing or measurement practice is performed, it can benefit from what has been presented above—be it in language and linguistics, psychology, behavioral sciences, or elsewhere. All in all, testing and assessment need to be precise, and guaranteeing precision needs complicated analytical methods. This paper tried to shed light on some of these analytical methods in measurement, but what was presented above is just a primer, and the interested reader is invited to read the sources that have been listed in the list of references below.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). John Wiley & Sons.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Brennan, R. L. (1984). Estimating the dependability of scores. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 292-334). The Johns Hopkins University Press.

Brown, J. D., & Salmani Nodoushan, M. A. (2015). Language testing: The state of the art (An online interview with James Dean Brown). *International Journal of Language Studies, 9*(4), 133-143.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues & Practice, 17*, 31-44. doi: 10.1111/j.1745-3992.1998.tb00619.x

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*, 269-279. doi: 10.1207/s15324818ame0604_2

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. doi: 10.1037/0033-2909.112.1.155

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 292-334.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). Harper and Row.

Cronbach, L. J., Geleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurement: Theory of generalizability for scores and profiles*. John Wiley & Sons.

de Ayala, R. J. (2009). The theory and practice of item response theory. Guilford.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18*, 131-154. doi: 10.2307/1165084

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407-424. doi: 10.1007/BF02288803

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (pp. 105-146). American Council on Education and Macmillan.

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling, 18*, 229-252. doi: 10.1080/10705511.2011.557338

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Bond.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (pp. 65-110). American Council on Education/Praeger.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer Academic Publishers.

Harris, D. P. (1969). *Testing English as a second language*. McGraw Hill.

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

Holmes Finch, W., Immekus, J. C., & French, B. F. (2016). *Applied psychometrics using SPSS and AMOS*. Information Age Publishing, Inc.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349. doi: 10.1207/S15324818AME1404_2

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125-160. doi: 10.1177/014662168200600201

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4*, 219-240.

Karami, H. (2018). On the impact of differential item functioning on test fairness: A Rasch modeling approach. *International Journal of Language Studies, 12*(3), 1-14.

Karami, H., & Salmani Nodoushan, M. A. (2011). Differential Item Functioning (DIF): Current problems and future directions. *International Journal of Language Studies, 5*(3), 133-142.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Houghton Mifflin.

Linn, R. L. (2009). The concept of validity in the context of NCLB. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 195-212). JAM Press.

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement, 22*, 511-521. doi: 10.1177/001316445701700407

Mantel, N., & W. Haenszel (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748. doi: 10.1093/jnci/22.4.719

Messick, S. (1988). Validity. In L. R. Linn (Ed.), *Educational measurement* (pp. 13-103). American Council on Education/McMillan.

Michaelides, M. P. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment, Research, and Evaluation, 13*(7), online. http://pareonline.net/getvn.asp?v=13&n=7

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176. doi: 10.1177/014662169201600206

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274. doi: 10.1177/014662169602000306

Nunemacher, J. (1999). Asymptotes, cubic curves, and the projective plane. *Mathematics Magazine, 72*(3), 183-192. doi:10.2307/2690881

Osborne, J. W. (2003). Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *PARE: Practical Assessment, Research, and Evaluation, 8*(11), online. doi: 10.7275/0k9h-tq64

Pedhazur, E. J., & Pedhazur Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach*. Lawrence Erlbaum Associates.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502. doi: 10.1007/BF02294403.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* University of Chicago Press.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116. doi: 10.1177/014662169301700201

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230. doi: 10.1111/j.1745-3984.1996.tb00490.x

Salmani Nodoushan, M. A. (2009). Measurement theory in language testing: Past traditions and current trends. *Journal on Educational Psychology, 3*(2), 1-12.

Salmani Nodoushan, M. A. (2020). Language assessment: Lessons learnt from the existing literature. *International Journal of Language Studies, 14*(2), 135-146.

Salmani Nodoushan, M. A. (2021a). Test affordances or test function? Did we get Messick's message right? *International Journal of Language Studies, 15*(3), 127-140.

Salmani Nodoushan, M. A. (2021b). Washback or backwash? Revisiting the status quo of washback and test impact in EFL contexts. *Studies in English Language and Education, 8*(3), 869-884. DOI:10.24815/siele.v8i3.21406

Shavelson, R. J., Webb, N., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*, 922-932. doi: 10.1037/0003-066X.44.6.922

Shavelson, R., & Webb, N. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133-166. doi: 10.1111/j.2044-8317.1981.tb00625.x

Spearman, (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295. doi: 10.1111/j.2044-8295.1910.tb00206.x

Swaminathan, H., & Rogers, H. J., (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370. https://www.jstor.org/stable/1434855

Thomas, D. R., & Zumbo, B. D. (1996). Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational & Behavioral Statistics, 21,* 110-130. doi: 10.2307/1165213

Valette, R. M. (1977). *Modern language testing* (2nd. ed.). Harcourt College Publication.

van der Linden, W., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *PARE: Practical Assessment, Research, and Evaluation, 12*(3), 1-26. doi: 10.7275/mhqa-cd89

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-154). American Council on Education and Praeger.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) scores*. Directorate of Human Resources Research and Evaluation, department of National defense.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics*: *Psychometrics* (vol.26, pp. 45-79). Elsevier Science B.V.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement*. (ETA RR-12-08). http://www.ets.org/research/policy_research_reports/publications/report/2012/jevu