An Overview of the History and Methodological Aspects of Psychometrics
History and Methodological aspects of Psychometrics

Luis ANUNCIAÇÃO
Federal University of Rio de Janeiro
Institute of Psychology, Department of Psychometrics
Rio de Janeiro, Brazil
E-mail: luisfca@gmail.com
Scientific article

Abstract

INTRODUCTION: The use of psychometric tools such as tests or inventories comes with an agreement and acceptance that psychological characteristics, such as abilities, attitudes or personality traits, can be represented numerically and manipulated according to mathematical principles. Psychometrics and its close relation with statistics provides the scientific foundations and the standards that guide the development and use of psychological instruments, some of which are tests or inventories. This field has its own historic foundations and its particular analytical specificities and, while some are widely used analytical methods among psychologists and educational researchers, the history of psychometrics is either widely unknown or only partially known by these researchers or other students.

OBJECTIVES: With that being said, this paper provides a succinct review of the history of psychometrics and its methods. From a theoretical approach, this study explores and describes the Classical Test Theory (CTT) and the Item Response Theory (IRT) frameworks and its models to deal with questions such as validity and reliability. Different aspects that gravitate around the field, in addition to recent developments are also discussed, including Goodness-of-Fit and Differential Item Functioning and Differential Test Functioning.

CONCLUSIONS: This theoretical article helps to enhance the body of knowledge on psychometrics, it is especially addressed to social and educational researchers, and also contributes to training these scientists. To a lesser degree, the present article serves as a brief tutorial on the topic.

Keywords: Psychometrics, History, Classical Test Theory, Item Response Theory, Measurement.

Citation: Anunciação, L. An Overview of the History and Methodological Aspects of Psychometrics-History and Methodological aspects of Psychometrics. Journal for ReAttach Therapy and Developmental Diversities. https://doi.org/10.26407/2018jrtdd.1.6

Corresponding address:
Luis Anunciação
Federal University of Rio de Janeiro
Institute of Psychology, Department of Psychometrics
Rio de Janeiro – Brazil 22290-902
E-mail: luisfca@gmail.com

Introduction

When one decides to use a psychological instrument such as a questionnaire or a test, the decision comes with an inherent understanding and agreement that psychological characteristics, traits or abilities can be investigated in a systematic manner. Another agreement is made when one decides to analyse the data obtained by a psychological tool by summing up the scores or by using other mathematical methods. This latter attitude comes with a deep epistemological acceptance that psychological traits can be casted in numerical form for the underlying structure. Although these premises were already well known and documented in publications by the first psychologists, this paradigm was not entirely accepted by the scientific community until recently.

Discussions about the general utility or validity of psychometrics are still present in the mainstream academic debate. Some authors argue against the utility or validity of psychometrics for answering questions about the underlying processes that guide observed behaviors (Toomela, 2010), and others say that the quantitative approach led psychology into a "rigorous science" (Townsend, 2008, p. 270). Apart from this discussion, the growth in the use of statistical and psychometric methods in psychological, social and educational research has been growing in recent years and some concerns have been expressed because of its inadequate, superficial or misapplied use (Newbery, Petocz, & Newbery, 2010; Osborne, 2010).

The close relationship between statistics and psychology is well documented and with the formation of the Psychometric Society in 1935 by L.L. Thurstone, psychometrics is seen as a separate science that interfaces with mathematics and psychology. In a broader sense, psychometrics is defined as the area concerned with quantifying and analysing human differences and in a narrower sense it is concerned with evaluating the attributes of psychological tests and other measures used to assess variability in behaviour and then to link such variability to psychological phenomena and theoretical frameworks (Browne, 2000; Furr & Bacharach, 2008). More recently, psychometrics also aims to develop new methods of statistical analysis or the refinement of older techniques, which has been possible with the advancements in computer and software technologies.

The two disciplines of psychometrics and statistics have at least three points in common. Firstly, they use models to simplify and study reality; secondly, they are highly dependent on mathematics; and thirdly, both can be observed by its tools (e.g. statistical inference tests are provided by statistics and/or psychological instruments are provided by psychometrics) or by their theoretical framework, where researchers seek to build new models and paradigms through guidelines, empirical data and simulations.

Strictly speaking, psychological phenomena such as attention and extraversion are not directly observable, nor can they be measured directly. Because of that, they must be inferred from observations made on some behaviour that may be observed and is assumed to operationally represent the unobservable characteristic (or "variable") that is of interest. There are numerous synonyms in the literature when referring to non-directly observable psychological phenomena such as abilities, constructs, attributes, latent variables, factors or dimensions (Furr & Bacharach, 2008).

There are several avenues available when trying to assess psychological phenomena. Multimethod assessments such as interviews, direct observation, and self-reporting, as well as quantitative tools such as tests and scales are accessible to psychologists (Hilsenroth, Segal, & Hersen, 2003). However, from this group of methods the use of tests, inventories, scales, and other quantitative tools are seen as the best choices when one needs to accurately measure

psychological traits (Borsboom, Mellenbergh, & van Heerden, 2003; Craig, 2017; Marsman et al., 2018; Novick, 1980), as long as they are psychometrically adequate.

In line with this, the use of quantitative methods in psychology (and social sciences in general) has been increasing dramatically in the last decades (since 1980s), despite strong criticism and concern from different groups that disagree with this quantitative view (Cousineau, 2007). Paradoxically, this quantitative trend was only partially followed by academics and other students of psychology, which has led to the American Psychological Association creating a task force aiming to increase the number of quantitative psychologists and to improve the quantitative training among students.

With that being said, the aim of this article is to provide a succinct review of the history of psychometrics and its methods through important points of psychometrics. It is important to clarify that this review is not about examining all trends in psychometrics so that it is not exhaustive and has concentrated on describing and summarising the topics related to this thesis. Several other resources are relevant to the topic and some are listed in the references.

History of Psychometrics
The precise historical origins of psychometrics and the field of quantitative psychology are difficult to define. The same condition is found in statistics when trying to detail when statistics was incorporated into social sciences/humanities. However, it is possible to argue that the investigation into psychometrics has two starting points. The first one was concerned with discovering general laws relating the physical world to observable behaviour and the second one had the aim to explore and to test some hypotheses about the nature of individual differences by using psychological testing (Craig, 2017; Furr & Bacharach, 2008). When arranging events in their order of occurrence in time, James McKeen Cattell was the first psychologist to

write about psychometrics in 1886 with a thesis entitled "Psychometric Investigation", in which he studied what we now know today as the Stroop effect. At this time, Cattel was Wundt's student, but he was highly influenced by Francis Galton and his "Anthropometric Laboratory" which opened in London in 1884. As consequence of the interface between the two researchers, Cattell is also credited as the founder of the first laboratory developed to study psychometrics, which was established within the Cavendish Physics Laboratory at the University of Cambridge in 1887 (Cattell, 1928; Ferguson, 1990).

With this first laboratory, the field of psychometrics could differentiate from psychophysics and the major differences can be grouped as the following: 1) while psychophysics aimed to discover general sensory-perception laws (i.e. psychophysical functions), psychometrics was (is) concerned with studying differences between individuals; 2) the goal of psychophysics is to explore the fundamental relations of dependency between a physical stimulus and its psychological response, but the goal of psychometrics is to measure what we call latent variables, such as intelligence, attitudes, beliefs and personality; 3) the methods in psychophysics are based on experimental design where the same subject is observed over repeated conditions in a controlled experiment, but the majority of studies in psychometrics are observational when the measurement occurs without trying to affect the participants (Jones & Thissen, 2007).

Nowadays, graduate programs in Psychometrics are found in countries such as the United States and division 5 (Quantitative and Qualitative Methods) from the American Psychological Association (APA) helps in studying measurement, statistics, and psychometrics. As can be captured in the definition of psychometrics, one of the primary strengths of psychometrics is to improve psychological science by developing instruments based on different theories and

approaches, thus, comparing its results. However, these instruments can be developed by other needs and areas (e.g. health sciences and business administration), which means that psychometric tools span across a variety of different disciplines.

With that being said, the Classical Test Theory (CTT) and the Item Response Theory (IRT) are the primary measurement theories employed by researchers in order to construct psychological assessment instruments and will be described in the following section.

Classical Test Theory (CTT) and Item Response Theory (IRT)

As there is no universal unit of psychological processes (or it has not been discovered *yet*), such as meters (m) or seconds (s), psychologists operate on the assumption that the units are implicitly created by the instrument that is being used in research (Rakover, 2012; Rakover & Cahlon, 2001). Two consequences emerge from this: first, there are several instruments to measure (sometimes the same) psychological phenomena; second, evaluating the attributes of psychological testing is one of the greatest concerns of psychometrics.

The indirect nature of the instruments leaves much room for unknown sources of variance to contribute to participant's results, which translates into a large measurement error and the conclusion that assessing the validity and the reliability of the psychometric instruments is vital (Peters, 2014). Additionally, as the data yielded by those tests are often used to make important decisions, including awarding credentials, judging the effectiveness of interventions and making personal or business decisions, ensuring that psychometric qualities remain up to date is a central objective in psychometrics (Osborne, 2010).

There are two distinct approaches/paradigms in psychometrics used in evaluating the quality of tests: CTT and IRT. Both deal with broad concepts such as validity, reliability and usability, and provide the mathematical guidance

to check test properties, as well as the epistemological background to address typical questions that emerge in psychometric research.

Validity is an extensive concept and has been widely debated since it was conceived in the 1920s. Throughout its history, at least three different approaches emerged to define it. The first authors had the understanding that validity was a test property (e.g. Giles Murrel Ruch, 1924; or Truman L. Kelley, 1927); the second conceived validity within a nomological framework (e.g. Cronbach & Meehl, 1955), and finally, current authors state that validity must not only consider the interpretations and actions based on test scores, but also the ethical consequences and social considerations (e.g. Messick, 1989) (Borsboom, Mellenbergh, & van Heerden, 2004).

There is no difficulty in recognising that the latter approach influenced official guidelines, such as the Standards of Testing, when it defines validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014, p. 14). However, even considering that psychometric tools always exist in a larger context and thus must be evaluated within this standpoint, this definition imposes a validation process which is hard to achieve. The absence of standard guidance for how to integrate different pieces of validity, or which evidence should be highlighted and prioritised contributes even more to weaken the link between theoretical understanding about validity and the practical actions performed by psychometricians to validate a tool (Wolming & Wikström, 2010).

Another effect of plural definitions is that not everyone has access to updated materials. This is pretty common in some cultures, mainly in developing countries, in which only translated content is available. Moreover, the types of validity elaborated by Cronbach and Meehl (1955) are not only older than the recent definitions, which increases its chances to have been translated, but are still informative and

reported in academic books. This mix between absence of updated knowledge about psychometrics and multiple ways to define the same concept nurtures an environment where analysis and conclusions can be diametrically opposed from one academic group to another.

Within this traditional framework, validity can be divided into content, criterion and construct (i.e. the "tripartite" perspective). Criterion-related validity is formed by concurrent and predictive validity. The construct-related validity is formed by convergent and discriminant validity. Finally, content refers to the degree an instrument measures all of the domains that constitute the domain and it is mainly assessed by experts in the domain. The statistical methods were developed or used for focusing on some particular aspect of validity, seen as independent of one another. However, as construct validity points to the degree to which an instrument measures what it is intended to measure, this type of validity became the central issue on the study of psychometrics (Cronbach & Meehl, 1955).

Nowadays, the progress of construct validity is accepted by virtually all psychometricians, in addition to the agreement that validity is not all or one nor a test property. Test validity should be evaluated within multiple sources of evidence with respect to specific contexts and purposes. Thus, the validation is a continuous process and a test can be valid for one purpose, but not for another (Sireci, 2007).

CTT is based on the concept of the "true score". That means the observed test score (Y_i) as composed of a True score (T_i) plus an Error (ϵ_i) considered normally distributed with its mean taken to be 0. The mathematical formulation of CTT have been made over the years until the work of Novick (1966), that defined:

$$Y_i = T_i + \epsilon_i$$

Equation 1. Basic CTT Equation

CTT accesses validity mainly by inter-item correlations, factor analysis and correlation between the measure and some external evidence (Salzberger, Sarstedt, & Diamantopoulos, 2016). CTT also understands reliability as a necessary but not sufficient condition for validity, while the reliability represents the consistency and the reproducibility of the results across different test situations.

To a lesser degree with what occurs for validity, this concept also has multiple meanings. It refers to at least three different concepts, which are internal consistency, consistency across time, and equivalence. Internal consistency is also referred to as item homogeneity and attempts to check if all the items of a test are relatively similar. Consistency across time is also known as temporal stability and is checked by consecutive measures of the same group of participants. Equivalence refers to the degree to which equivalent forms of an instrument or different raters yield similar or identical scores (AERA, APA, & NCME, 2014; Borsboom et al., 2004; Sijtsma, 2013).

From a CTT perspective, reliability is the ratio of true-variance to the total variance yielded by the measuring instrument. The variance is:

$$I?_Y^2 = I?_T^2 + I?_E^2$$

Equation 2. Decomposition of the test variances

Hence, the reliability is:

$$\rho^2_{(YT)} = \frac{\sigma^2_T}{\sigma^2_Y} \therefore \rho^2_{(YT)} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E}$$

Equation 2. Decomposition of the test reliability

As reliability is not a unitary concept, several methods were developed for its evaluation such as Cronbach's alpha, Test-retest, Intraclass Correlation Coefficient (ICC) and Pearson or Spearman correlation. Cronbach's alpha is the most commonly used to measure the internal consistency and has proven to be very resistant to the passage of time, despite its limitations: its values are dependent on the number of items in the scale, assumes tau-equivalence (i.e. all factor loadings are equal or the same true score for all test items), is not robust against missing data, and treats the items as continuous and normally distributed data (McNeish, 2017). Alternatives to Cronbach's alpha have been proposed, and examples are the McDonald's omega, The Greatest Lower Bound (GLB) and Composite Reliability (Sijtsma, 2009).

Still within the CTT framework, a shift has occurred with Factor Analysis (FA). This method relies on a linear model and depends on the items included in the test and the persons examined, but it also models a latent variable and some of its models achieve virtually identical results to those obtained by IRT models. Therefore, these conditions allow that one considers FA from both perspectives/traditions in psychometrics (Steyer, 2001). If the framework in this article uses the "true vs latent variable", FA will be allocated into latent framework such as IRT, and from a statistical/methodological standpoint it is possible to combine approaches or understand some methods as particular cases of a general approach, such as with Confirmatory Factor Analysis (Edwards & Bagozzi, 2000; Mellenbergh, 1994).

In regards to the statistical process to explore the constructs covered in psychometric work, there are two main ways in which this connection between constructs and observations has been construed. The first approach understands constructs as inductive summaries of attributes or behaviours as a function of the observed variables (i.e. formative model, where latent variables are formed by their indicators). The second approach understands constructs as reflective and the presence of the construct is assumed to be the common cause of the observed variables (Fried, 2017; Schmittmann et al., 2013). Image 1 below displays these conceptualisations.
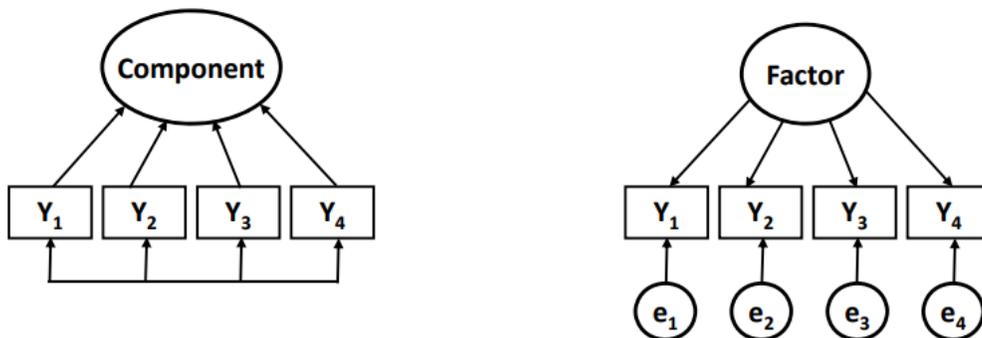

Image 1. On the left, the PCA model (formative); on the right, the Factor model (reflective).

As the goal of PCA is data reduction, but psychometric theory wants to investigate how

observable variables are related to theoretical/latent constructs, the reflective model

is mostly used. Some of the statistical models associated with this model are the Common Factor Model, Item Response Theory models (IRT), Latent Class Models, and Latent Profile Models (Edwards & Bagozzi, 2000; Marsman et al., 2018). The question whether latent variables are continuous (therefore dimensions) or categorical (therefore typologies) will influence the choice of the model. Table 1 reports the theoretical assumptions of latent and manifest variables in reflective models.

Table 1. Properties of Latent and Observed variables

Model	Latent variable (ability, trait)	Observed variable (items, indicators)
Common Factor	Continuous	Continuous
Item Response Theory	Continuous	Categorical
Latent Class Analysis	Categorical	Categorical
Latent Profile Analysis	Categorical	Continuous

The Factor Analysis (FA) is part of its models, its concept is analogous to CTT and was developed with the work of Charles Spearman (1904) in the context of intelligence testing. The FA operates on the notion that measurable and observable variables can be reduced to fewer latent variables that share a common variance and are unobservable (Borsboom et al., 2003). The statistical purpose of factor analysis is to explain relations among a large set of observed variables using a small number of latent/unobserved variables called factors. FA can be divided into exploratory and confirmatory, and in a broad sense is viewed as a special case of Structural Equation Modeling (SEM) (Gunzler & Morris, 2015).

Exploratory factor analysis (EFA) explores data to determine the number or nature of factors that account for the covariation between variables if the researcher does not own sufficient *a priori* evidence to establish a hypothesis regarding the number of factors underlying the data. In detail, since there is not an a priori hypothesis about how indicators are related to the underlying factors, EFA is not generally considered a member of the SEM family. In contrast, confirmatory factor analysis (CFA) is a theory-driven model and aims to see whether a particular set of factors can account for the correlations by imposing lower triangular constraints on the factor loading matrix, thus rendering identifiability to the established parameters of the model. In other words, CFA is designed to evaluate the a priori factor structure specified by researchers (Brown, 2015; Finch, 2011).

In another direction, some authors argue that there is no clear EFA-CFA distinction in most factor analysis applications and they fall on a continuum running from exploration to confirmation. Because of this, they choose to call both techniques at a statistics standpoint; an unrestricted model for EFA and a restricted model for CFA. An unrestricted solution does not restrict the factor space, so unrestricted solutions can be obtained by a rotation of an arbitrary orthogonal solution and all the unrestricted solutions will yield the same fit for the same data. On the other hand, a restricted solution imposes restrictions on the whole factor

space and cannot be obtained by a rotation of an unrestricted solution (Ferrando & Lorenzo-Seva, 2000).

Leaving aside these particular questions, several high-quality resources on best practices in EFA and CFA are available, and despite some changes in the mathematical notation or formula, the common factor model is a linear regression model with observed variables as outcomes (dependent variables) and factors as predictors (independent variable) (see equation 1):

$$Y_i = \left(\sum_{m=1}^{M} \lambda_{im} \eta_m \right) + \epsilon_i$$

Equation 3. Common factor model

Where Y_i is the i*th* observed variable (item score) from a set of I observed variables, η_m is the m*th* of M common factors, λ_{im} is the regression coefficient (slope, also known as factor loading) relating factor m to Y_i, and ϵ_i is the error term unique for each Y_i. The variance of ε for variable *i* is known as the variable's uniqueness, whereas 1 – VAR(ϵ) is that variable's communality. This latter concept is equivalent to the regression R^2 and describes the proportion of variability in the observed variable explained by the common factors. In some guidelines, the inclusion of the item intercept $I?_i$ is made, but this parameter usually does not contribute to the covariance matrix (Furr & Bacharach, 2008).

Operationally, some assumptions must be fulfilled before an EFA, such as the proportion of variance among variables that might be common variance, and that the dependent variable covariance matrices are not equal across the levels of the independent variables. The first assumption is tested by the Kaiser-Meyer-Olkin (KMO) test and the second with the Bartlett test. KMO values between 0.8 and 1 indicate the data is adequate for FA, and a significant Bartlett's test (p < .05) means that data matrix is not an identity matrix, which prevents factor analysis from working (Costello & Osborne, 2011).

Next, three main questions arise when conducting an EFA: 1. The method of factor extraction; 2. How many factors to settle on for a confirmatory step; and 3. Which factor rotation should be employed. All questions need to be answered by the researcher. The extraction methods reflect the analyst's assumptions about the obtained factors. Their mathematical conceptualisation is also based on manipulations of the correlation matrix to be analysed. There are a number of factors to retain changes throughout the literature and there are many rules of thumb to guide the decision. Finally, all results are often adjusted to become more interpretable.

In summary, the factor extraction methods are statistical algorithms used to estimate loadings and are composed of techniques such as the minimum residual method, principal axis factoring, weighted least squares, generalized least squares and maximum likelihood factor analysis. The decision of how many factors will be retained relies on many recommendations such as: 1. The rule of an eigenvalue of ≥ 1; 2. The point in a scree plot where the slope of the curve is clearly leveling off; or 3. The interpretability of the factors. It is easy to recognise that these guides can provide contradictory answers and illustrate some degree of arbitrary decisions during this process (Nowakowska, 1983). The factor rotations are classified as either orthogonal, in which the factors are constrained to be uncorrelated (e.g. Varimax, Quartimax, Equamax), or oblique (e.g. Oblimin, Promax, Quartimin) in which this constraint is not present (Finch, 2011).

Another approach in psychometrics independent of the factor analysis developments and apart from CTT is the IRT. The focus of IRT modeling

is on the relation between a latent trait (θ_s) , the properties of each item in the instrument and the individual's response to each item. IRT assumes that the underlying latent dimension (or dimensions) are causal to the observed responses to the items, and different from CTT, item and person parameters are invariant, neither depending on the subset of items used nor on the distribution of latent traits in the population of respondents. In addition, the total scores of a test has no space in IRT, which is concerned with focusing on quality at the item level.

Considering a sample of *n* individuals that answered *I* items. *s* = 1, …, *n* and *i* = 1, ..., *I*. Let Y_{ij} be random variables associated with the

response of individual *s* to an item *i*. These responses can be dichotomous (e.g. fail or pass) or polytomous (e.g. agree, partially agree, neutral). Let Ω_Y denote the set of possible values of the Y_{ij}, assumed to be identical for each item in the test, and θ_s denotes the latent trait for an individual *s*, and η_i a set of parameters that will be used to model item features. The IRT models arise from different sets of possible responses Ω_Y and different functional forms assumed to describe the probabilities with which the Y_{ij} assume those values, as expressed below (Le, 2014; Sijtsma & Junker, 2006; Zumbo & Hubley, 2017):

$$P(Y_{ij} = y|\theta_s, \eta_i) = f(y|\theta_s, \eta_i); y \in \Omega_Y$$

Equation 4. General formula of IRT models

The η_i represents the item parameters and may include four distinct types of parameters: parameter "a_i" denotes the discrimination, "b_i" the difficulty, "c_i" the guessing, and "d_i"

expresses the probability of a high-ability participant failing to answer an item correctly. The common 4PL model for a dichotomous response is (Loken & Rulison, 2010):

$$P(Y_{is} = 1 | \theta_s, a, b, c, d) = c_i + (d_i - c_i) \; e^{[a_i(\theta_s - b_i)]} / (1 + e^{([a_i(\theta_s - b_i)])}), a_i > 0; 0 \le c \le 1$$

Equation 5. 4PL IRT model

Which leads to:

$$P(Y_{is} = 1 | \theta_s, a, b, c, d) = c_i + (d_i - c_i) \; 1/(1 + e^{([-a_i(\theta_s - b_i)])})$$

Equation 6. 4he concept PL IRT model

The three IRT models that precede the 4PL are seen as its constrained version. The 3PL model constrains the upper asymptote ("d") to 1, the 2PL model keeps the previous constraint and also constrains the lower asymptote ("c") to 0, and the 1PL model only estimates the difficulty parameter ("b"). Some information about these models must be emphasised for better understanding of the topic: 1. The 2PL is analogous to the congeneric measurement model in CTT, 2. Both the 1PL and Rasch models assume that items do not differ in the

discrimination parameter ("a"), but Rasch models set the discrimination at 1.0, whereas 1PL can assume other values, and 3. Some authors argue that Rasch models focus on fundamental measurement, trying to check how well the data fits the model, while IRT models check the degree to which the model fits the data (De Ayala, 2009, p. 19).

As can be seen from the equations, there is a conceptual bridge between IRT parameters and Factor Analysis, and between IRT models and logistic regression. The "a" parameter is

analogous to the factor loading in traditional linear factor analysis, with the advantage that the IRT model can accommodate items with different difficulties, whereas linear factor loadings and item-total correlations will treat easy or hard items as inadequate because they have less variance than medium items. The "b" parameter in Rasch models is analogous to item difficulty in CTT, which is the probability of answering the item correctly (Schweizer & DiStefano, 2016).

Similarities also exist between IRT and logistic regression, but the explanatory (independent) variable in IRT is a latent variable as opposed to an observed variable in logistic regression. In the IRT case, the model will recognize the person's variability on the dimension measured in common by the items and individual differences θ may be estimated (Wu & Zumbo, 2007).

In the origins of IRT, some assumptions (such as unidimensionality and local independence) were held, but IRT models can currently deal with multidimensional latent structure (MIRT) and local dependence. In MIRT, an Item Characteristic Surface (ICS) represents the probability that an examinee with a given ability (θ_s) composite will correctly answer an item. To deal with local independence, Item Splitting is a way for the estimation of item and person parameters (Olsbjerg & Christensen, 2015). In the same direction, the comparison between unidimensional and multidimensional models have shown that as the number of latent traits underlying item performance increase, item and ability parameters estimated under MIRT have less error scores and reach more precise measurement (Kose & Demirtasli, 2012).

As previously stated, the reliability of an instrument is investigated along with the validity during a psychometric examination of an instrument, and it can be performed via methods within the CTT and IRT framework. In IRT, reliability varies for different levels of the latent trait, meaning that the items

discriminate better around their difficulty parameter.

It should be emphasised that both CTT and IRT methods are currently seen as complementary and are frequently used to assess the test validity and respond to other research questions.

Goodness-of-fit (GoF)

As most modern measurement techniques do not measure the variable of interest directly, but indirectly derive the target variables into models, the adequacy of models must be tested by statistical techniques and experimental or empirical inspection. Goodness-of-Fit (GoF) is an important procedure to test how well a model fits a set of observations or whether the model could have generated the observed data. Both SEM and IRT provide a wide range of GoF indices focusing on the item and/or test-level, and the guidelines in SEM are seen as reasonable for IRT models (Maydeu-Olivares & Joe, 2014).

Traditional GoF indices can be broken into absolute and relative fit indices. The absolute measures the discrepancy between a statistical model and the data, whereas the relative measures the discrepancy between two statistical models. The first indices are comprised of Chi-Square, Goodness of Fit Index (GFI), Root Mean Square of Approximation (RMSEA), and Standardized Root Mean-Square Residual (SRMR). The second indices are comprised of Bollen's Incremental Fit Index (IFI), Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI).

Mainly in Rasch-based methods, some item-level fit indices are also available to assess the degree to which an estimated item response function approximates (or does not) an observed item response pattern. Finally, the information-theoretic approach is a commonly used criteria in model selection, with the Akaike information criterion (AIC) and the Bayesian information (BIC) being the most used measures to select from among several candidate models (Fabozzi,

Focardi, Rachev, & Arshanapalli, 2014). Table 2 summarizes these indices; it may be used as a preliminary approach to these models and is based on Bentler (1990), Maydeu-Olivares (2013), Fabozzi et al. (2014), and Wright & Linacre (1994).

Table 2. Measures of Goodness-of-Fit

Commonly used Framework	Type	Indices	Values
SEM	Absolute	Chi-Square (χ^2) RMSEA SRMR	P value ≥ 0.05 ≤ 0.08 – acceptable ≤ 0.05 – ideally
	Relative	TLI CLI ILI	≥ 0.95 – ideally ≥ 0.90 – acceptable
Rasch/IRT	Item level	Infit Outfit	Based on type of test, for surveys, 0.6 - 1.4
Information Criteria	Methods for comparing competing models	AIC BIC	Lowest value Lowest value

Differential item functioning (DIF) and Differential Test Functioning (DTF)

DIF refers to the change in the probability of participants within the same ability level, but from different groups, in successfully answering a specific item. Therefore, assuming two individuals from different subgroups have the same ability level, their probability of endorsing the same items should not be different. When DIF is present for many items on the test, the final test scores do not represent the same measurement across groups, and this is known as DFT (Runnels, 2013).

DIF (and DFT) may reflect measurement bias and indicate a violation of the invariance assumption. Testing DIF is enabled by visual inspection and statistical testing. As the Item Characteristic Curves represents the regression of the item score (dependent variable) on examinees' ability, different patterns emerging from groups with the same ability is the first evidence of DIF. In addition, Mantel-Haenszel (MH), Wald statistics, and the Likelihood-ratio test approach offer a numerical approach to investigate DIF (De Beer, 2004).

As meaningful comparisons require that measurement equivalence holds, both DIF and DFT may influence the psychometric properties of test scores and represent lack of fairness. Further literature about DIF and DFT are available elsewhere (Hagquist & Andrich, 2017).

Conclusions

The investigator often needs to simplify some representation of reality in order to achieve an understanding of the dominant aspects of the system under study. This is no different in Psychology; models are built and their study

allows researchers to answer focused and well-posed questions. When models are useful, their predictions are analogous to the real world. Additionally, psychometric tools are often used to inform important decisions, judging the effectiveness of interventions, and making personal or business decisions. Therefore, ensuring that psychometric qualities remain up to date is a central objective in psychometrics (Osborne, 2010).

The present manuscript had the goal to explore some aspects of the history of psychometrics and to describe its main models. Theoretical studies frequently focus on one of the two aspects. However, the integration of methods and its history helps to better understand (and contextualise) psychometrics. The preceding pages revisited the origins of psychometrics through its models, as well as illustrated some of the mathematical conceptualisations of these techniques, in addition to academic perspectives on psychometrics.

Despite the contributions provided in this manuscript, it is not free from limitations. The present text does not cover some of the recent methods and debate, such as Bayesian psychometrics, network psychometrics and the effect of computational psychometrics on psychology. Bayesian psychometrics in particular, and Bayesian statistics in general are seen as candidates to make a revolution in Psychology and other behavioural sciences (Kruschke, Aguinis, & Joo, 2012). Along these same lines, the potential candidates to change current psychometric paradigms are network models. Different from the traditional view that understands item response being caused by latent variables, in network models items are hypothesised to form networks of mutually reinforcing variables (Fried, 2017). Finally, the growth in computer power and the availability of statistical packages can negatively impact psychometrics by encouraging a generation of mindless analysts if uncorrelated with the theoretical understanding of science and the scientific method.

Because the use of psychometric tools is becoming an important part of several sciences, understanding the concepts presented in this paper will mainly be of importance to enhance the abilities of social and educational researchers.

Acknowledgements

Conflicts of interests

The author declares no conflict of interests.

References

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. In *American Educational Research Association*.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brown, T. (2015). *Confirmatory Factor Analysis for Applied Reaearch. Journal of Chemical Information and Modeling*. https://doi.org/10.1017/CBO9781107415324.004

Browne, M. W. (2000). Psychometrics. *Journal of the American Statistical Association*, *95*(450), 661–665. https://doi.org/10.1080/01621459.2000.10474246

Cattell, J. M. (1928). Early Psychological Laboratories. *Science*, *67*(1744), 543–548. https://doi.org/10.1126/science.67.1744.543

Costello, A., & Osborne, J. (2011). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*, np.

Cousineau, D. (2007). The rise of quantitative methods in psychology. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 1–3. https://doi.org/10.20982/tqmp.01.1.p001

Craig, K. (2017). The History of Psychometrics. In *Psychometric Testing* (pp. 1–14). Wiley. https://doi.org/10.1002/9781119183020.ch1

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0061470

De Ayala, R. J. (2009). The Theory and Practice of Item Response Theory. *Methodology in the Social Sciences*.

De Beer, M. (2004). Use of differential item functioning (DIF) analysis for bias analysis in test construction. *SA Journal of Industrial Psychology*, *30*(4). https://doi.org/10.4102/sajip.v30i4.175

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Fabozzi, F. J., Focardi, S. M., Rachev, S. T., & Arshanapalli, B. G. (2014). Multiple Linear Regression. In *The Basics of Financial Econometrics* (pp. 41–80). https://doi.org/10.1002/9781118856406.ch3

Ferguson, B. (1990). Modern psychometrics. The science of psychological assessment. *Journal of Psychosomatic Research*, *34*(5), 598. https://doi.org/10.1016/0022-3999(90)90043-4

Ferrando, P. J., & Lorenzo-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: some aspects of the problem and some suggestions. *Psicológica*, *21*, 301–323.

Finch, W. H. (2011). A Comparison of Factor Rotation Methods for Dichotomous Data. *Journal of Modern Applied Statistical Methods*, *10*(2), 549–570. https://doi.org/10.22237/jmasm/1320120780

Fried, E. I. (2017). What are psychological

constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health Psychology Review*, *11*(2), 130–134. https://doi.org/10.1080/17437199.2017.130 6718

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics : an introduction*. Retrieved from http://catdir.loc.gov/catdir/enhancements/fy0 808/2007016663-b.html

Gunzler, D. D., & Morris, N. (2015). A tutorial on structural equation modeling for analysis of overlapping symptoms in co-occurring conditions using MPlus. *Statistics in Medicine*, *34*(24), 3246–3280. https://doi.org/10.1002/sim.6541

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health and Quality of Life Outcomes*, *15*(1), 181. https://doi.org/10.1186/s12955-017-0755-0

Hilsenroth, M. J., Segal, D. L., & Hersen, M. (2003). *Comprehensive handbook of psychological assessment. Assessment* (Vol. 2). https://doi.org/10.1002/9780471726753

Jones, L. V., & Thissen, D. (2007). A History and Overview of Psychometrics. *Handbook of Statistics*, *26*(6), 1–27. https://doi.org/10.1016/S0169-7161(06)26001-2

Kose, I. A., & Demirtasli, N. C. (2012). Comparison of Unidimensional and Multidimensional Models Based on Item Response Theory in Terms of Both Variables of Test Length and Sample Size. *Procedia - Social and Behavioral Sciences*, *46*, 135–140. https://doi.org/10.1016/j.sbspro.2012.05.082

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time Has Come. *Organizational Research Methods*, *15*(4), 722–752. https://doi.org/10.1177/1094428112457829

Le, D.-T. (2014). Applying item response theory modeling in educational research.

Dissertation Abstracts International: Section B: The Sciences and Engineering, *75*(1).

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509–525. https://doi.org/10.1348/000711009X474502

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., … Maris, G. (2018). An Introduction to Network Psychometrics: Relating Ising Network Models to Item Response Theory Models. *Multivariate Behavioral Research*, *53*(1), 15–35. https://doi.org/10.1080/00273171.2017.137 9379

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement: Interdisciplinary Research & Perspective*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831 680

Maydeu-Olivares, A., & Joe, H. (2014). Assessing Approximate Fit in Categorical Data Analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911 075

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*. https://doi.org/10.1037/met0000144

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*(2), 300–307. https://doi.org/10.1037/0033-2909.115.2.300

Newbery, G., Petocz, A., & Newbery, G. (2010). On Conceptual Analysis as the Primary Qualitative Approach to Statistics Education Research in Psychology. *Statistics Education Research Journal*, *9*(2), 123–145.

Novick, M. R. (1966). The axioms and principal results of classical test theory.

Journal of Mathematical Psychology, *3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

Novick, M. R. (1980). Statistics as psychometrics. *Psychometrika*, *45*(4), 411–424. https://doi.org/10.1007/BF02293605

Nowakowska, M. (1983). Chapter 2 Factor Analysis: Arbitrary Decisions within Mathematical Model (pp. 115–182). https://doi.org/10.1016/S0166-4115(08)62370-5

Olsbjerg, M., & Christensen, K. B. (2015). Modeling local dependence in longitudinal IRT models. *Behavior Research Methods*, *47*(4), 1413–1424. https://doi.org/10.3758/s13428-014-0553-0

Osborne, J. W. (2010). Challenges for quantitative psychology and measurement in the 21st century. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2010.00001

Peters, G.-J. Y. (2014). The alpha and the omega of scale reliability and validity. *The European Health Psychologist*, *16*(2), 56–69.
https://doi.org/10.17605/OSF.IO/TNRXV

Rakover, S. S. (2012). Psychology as an Associational Science: A Methodological Viewpoint. *Open Journal of Philosophy*, *2*(2), 143–152. https://doi.org/10.4236/ojpp.2012.22023

Rakover, S. S., & Cahlon, B. (2001). *Face Recognition* (Vol. 31). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/aicr.31

Runnels, J. (2013). Measuring differential item and test functioning across academic disciplines. *Language Testing in Asia*, *3*(1), 9. https://doi.org/10.1186/2229-0443-3-9

Salzberger, T., Sarstedt, M., & Diamantopoulos, A. (2016). Measurement in the social sciences: where C-OAR-SE delivers and where it does not. *European Journal of Marketing*, *50*(11), 1942–1952. https://doi.org/10.1108/EJM-10-2016-0547

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., &

Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, *31*(1), 43–53. https://doi.org/10.1016/j.newideapsych.2011.02.007

Schweizer, K., & DiStefano, C. (2016). *Principles and Methods of Test Construction: Standards and Recent Advances*. Hogrefe Publishing. https://doi.org/10.1027/00449-000

Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K. (2013). Theory Development as a Precursor for Test Validity (pp. 267–274). Springer New York. https://doi.org/10.1007/978-1-4614-9348-8_17

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika*, *33*(1), 75–102. https://doi.org/10.2333/bhmk.33.75

Sireci, S. G. (2007). On Validity Theory and Test Validation. *Educational Researcher*, *36*(8), 477–481. https://doi.org/10.3102/0013189X07311609

Steyer, R. (2001). Classical (Psychometric) Test Theory. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 1955–1962). Elsevier. https://doi.org/10.1016/B0-08-043076-7/00721-X

Toomela. (2010). Quantitative methods in psychology: inevitable and useless. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2010.00029

Townsend, J. T. (2008). Mathematical psychology: Prospects for the 21st century: A guest editorial. *Journal of Mathematical Psychology*, *52*(5), 269–280. https://doi.org/10.1016/j.jmp.2008.05.001

Wolming, S., & Wikström, C. (2010). The concept of validity in theory and practice.

Assessment in Education: Principles, *Policy and Practice*, 17(2), 117-132.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and Investigating Response Processes in Validation Research* (Vol. 69). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-56129-5