Papeles del Psicólogo, 2010. Vol. 31(1), pp. 57-66
http://www.cop.es/papeles

TEST THEORIES: CLASSICAL THEORY AND ITEM RESPONSE THEORY

José Muñiz
Psychology Faculty, University of Oviedo

For the correct interpretation and proper use of the psychometric properties of tests it is necessary to go beyond mere empirical calculation to a knowledge of the foundations on which these calculations are based. In an effort to contribute to this understanding beyond the superficial handling of the psychometric formulas, the main goal of this work is to present, in a non-technical way, the two most important theories that guide the development and analysis of most tests: Classical Test Theory and Item Response Theory. First, we offer a historical overview of tests and testing, describing how tests evolved in line with technical and statistical advances. The importance of test theories in the development and analysis of tests is discussed, and Classical Test Theory, including aspects such as Generalizability Theory and Criterion-Referenced Tests, is presented. After highlighting the limitations of the Classical Test Theory approach, Item Response Theory is presented. Within this new framework some of the limitations of the Classical Test Theory find a more satisfactory solution. Finally, the two approaches are compared, and the importance of test theories for the correct use and interpretation of the psychometric properties of tests is emphasized.

Key words: *Tests, Classical Test Theory, Item Response Theory, Test theories.*

Para una interpretación y utilización adecuada de las propiedades psicométricas de los tests es necesario ir más allá del mero cálculo empírico, y conocer los fundamentos en los que se basan esos cálculos. Con el fin de contribuir a esta comprensión más allá del mero manejo superficial de la fórmulas psicométricas, el objetivo fundamental de este trabajo es presentar de una manera no excesivamente técnica y especializada las dos grandes teorías que guían la construcción y análisis de la mayoría de los tests: la Teoría Clásica de los Tests y la Teoría de Respuesta a los Ítems. En primer lugar se hace un apunte histórico sobre los tests, indicando cómo surgen y evolucionan al hilo de los avances técnicos y estadísticos. Tras razonar acerca de la necesidad de utilizar teorías psicométricas para el análisis y construcción de los tests, se expone la lógica que subyace a la Teoría Clásica de los Tests, así como sus dos variantes más granadas, la Teoría de la Generalizabilidad y los Tests Referidos al Criterio. Luego se subrayan las limitaciones más importantes del enfoque clásico y se exponen los fundamentos de la Teoría de Respuesta a los Ítems, dentro de cuyo marco encuentran una solución satisfactoria algunos de los problemas que el enfoque clásico no había sido capaz de resolver de forma satisfactoria. Finalmente se comparan ambos enfoques, y se concluye indicando la necesidad de conocer las teorías de los tests para una mejor comprensión y utilización de los instrumentos de medida.

Palabras clave: *Tests, Teoría Clásica de los Tests, Teoría de Respuesta a los Ítems, Teorías de los tests.*

Tests undoubtedly constitute the most sophisticated technology available to psychologists in the exercise of their profession, so that it is not uncommon for society to identify psychologists with tests. Of course, some psychologists use tests more than others, depending on their professional field and way of working. Tests are samples of behaviour that permit us to make relevant inferences about people's behaviour. Used properly, they are key tools in the psychologists' profession. It should be borne in mind that tests have emerged out of a desire for objectivity and justice, to assess people on the basis of their true worth, avoiding evaluations biased by aspects such as background, social class, race, sex, beliefs or letters of recommendation, or by other subjective assessment systems. Such noble aims have been achieved more successfully in some cases than in others, but the central idea always has been, and continues to be that of assessing everyone based on the same criteria.

HISTORICAL OVERVIEW

When do tests appear for the first time? Often cited in reference to the remote origins of tests are the exams first used by Chinese emperors some 3000 years before Christ to assess the professional competence of their civil servants. Many other ancient precursors to the test can also be found, but the tests of today are most clearly descended from the sensory-motor tests used by Galton

Correspondence: José Muñiz. *Facultad de Psicología. Universidad de Oviedo. Plaza Feijoo, s/n. 33003 Oviedo. España. E-mail: jmuniz@uniovi.es*

(1822-1911) in his anthropometric laboratory. However, it would be James McKeen Cattell (1860-1944) who first used the term mental test, in 1890. It soon became clear (Wissler, 1901) that these first sensory-motor tests were not good predictors of people's cognoscitive capacities, and Binet and Simon (1905) would take things in a radically new direction on including in their new scales cognoscitive tasks to evaluate aspects such as judgement, comprehension and reasoning. Terman carried out his revision of the scale at Stanford University, for which reason it became known as the Stanford-Binet revision (Terman, 1916), using for the first time the concept of Intelligence Quotient (IQ) to express testees' scores. The idea of IQ had originally been proposed by Stern, who divided mental age by chronological age and multiplied the result by 100 to avoid decimals.

The Binet scale begins a tradition of individual scales that has continued up to the present day. In 1917 tests took another important leap forward with the emergence of the collective tests Alpha and Beta, borne out of the US Army's need for the rapid recruitment of men to serve in the First World War. The Alpha test was designed for the general population and the Beta test for those who were illiterate or without a good command of English. These tests were highly successful, and in the years following the Great War, companies and other institutions became enthusiastic in the use of them for various purposes. This marked the beginning of a boom in the use and creation of all types of tests. The emergence of the factor analysis technique represented a great advance in the construction and analysis of tests, paving the way for the appearance of test batteries, whose most genuine representative would be Thurstone's Primary Mental Abilities (PMA) test (Thurstone, 1938; Thurstone & Thurstone, 1941). In Spain we had the luck that one of the great pioneers of psychology in this country, Mariano Yela, studied in Chicago with Thurstone during the 1940s. As a result, he was able to introduce into Spain all the advances of the time, giving a boost to Psychometrics in both the academic world and applied contexts, and collaborating actively in the launching of the TEA publishing company (Pereña, 2007). The division of intelligence into its different factors or dimensions gave rise to the emergence of two broad lines of structuring of the cognoscitive dimensions, which have come to be known as the English school and the American school. The former gives more importance to a core factor of general intelligence, which would crown a structure involving two broad dimensions, verbal-educational and mechanical-spatial, in which many other more specific factors would be articulated. The American approach assumes a series of non-hierarchical dimensions that would make up the cognoscitive profile, which in the case of the PMA, for example, would be: verbal comprehension, verbal fluency, numerical ability, spatial ability, memory, perceptual speed and general reasoning. The two approaches are compatible, and have a good deal to do with the statistical technology employed, especially factor analysis. This whole line of psychometric research on intelligence culminates in Carroll's (1993) classic work, which synthesizes the great advances made. In Spain, works such as those of Juan-Espinosa (1997), Colom (1995) and Andrés-Pueyo (1996) brilliantly review and analyze this field.

But not only do advances take place in the field of cognoscitive tests: personality tests also take advantage of the progress being made in psychometrics. The personal data sheet used by Woodworth in 1917 for detecting severe neurotics is usually cited as the pioneer of personality tests. A little later, in 1921, the Swiss psychiatrist Rorschach proposed his projective inkblot test, which would be followed by many other tests based on the principle of projection, which assumes that faced with an ambiguous stimulus the person being assessed will tend to produce responses that in some way reflect important aspects of their personality. Readers interested in the history of tests may consult, for example, the work by Anastasi and Urbina (1998); here we offer only a brief outline to aid the understanding of what follows.

Now that tests have a century or so of history behind them, one might wonder which ones are the most widely used in Spain today, and if those commonly employed here differ from those which our colleagues in the rest of Europe tend to use. According to a recent survey carried out in six European countries, the tests most often used by Spanish psychologists were 16PF, WISC, WAIS, MMPI, Beck, STAI, Rorschach, Raven, Bender and ISRA, making the Spanish case similar to those of other countries in Europe (Muñiz et al., 2001).

In sum, the history of tests is a success story of which psychology should feel proud, not forgetting of course, that as with any technology in any field, sometimes their use by non-experts has left much to be desired. With this in mind, various organizations, both national (Spanish Psychological Association [*Colegio Oficial de Psicólogos*, COP]) and international (European Federation of

Psychologists' Associations, EFPA; International Test Commission, ITC; American Psychological Association, APA), have developed projects and activities to foster the appropriate use of tests (Muñiz, 1997b; Muñiz & Bartram, 2007; Prieto & Muñiz, 2000).

WHY DO WE NEED TEST THEORIES?

The previous section provided a brief historical outline of how specific tests emerged and evolved, but nothing has been said so far about the theories that make possible the construction of tests. Indeed, the reader may think that tests appear haphazardly, but nothing could be farther from the truth. Underlying the construction and analysis of tests are theories that guide their design and influence them according to the state of theoretical and statistical progress at the time.

But one might quite understandably ask oneself why we actually need test theories. Or in more pragmatic terms, why does the psychology degree course include the subject of Psychometrics, which is basically devoted to the exposition of these theories? The reason is quite simple: tests are sophisticated measurement instruments by means of which psychologists make inferences and decisions about important aspects of persons. Therefore, we must ensure that these inferences are appropriate and pertinent if we are to avoid our work being potentially detrimental to those who seek the help of a psychologist for any reason. Statistical theories of tests will permit the estimation of the psychometric properties of tests so as to guarantee that the decisions made on their basis are appropriate. Without these theories we could not estimate the reliability and validity of tests, and knowledge of these is essential to be able to use tests rigorously and scientifically. Of course, apart from these statistical theories of tests, the construction of a test must be guided by a substantive psychological model or theory. The work by Muñiz and Fonseca-Pedrero (2008) sets out the basic steps for constructing a test. For a more detailed analysis of the process of test construction the reader can consult, for example, the works by Carretero and Pérez (2005), Downing and Haladyna (2006), Morales, Urosa and Blanco (2003), Muñiz (2000), Schmeiser and Welch (2006), or Wilson (2005).

There are two principal approaches or theories in relation to the construction and analysis of tests, Classical Test Theory (CTT) and Item Response Theory (IRT). Here we shall not go into great detail about these theories (see, in Spanish, for example, Muñiz, 1997a, 2000, 2005), but rather highlight their key aspects, so that test users might gain a clearer idea and understand in more depth the implications of the psychometric properties of the tests they are using.

CLASSICAL TEST THEORY

The classical approach is that which predominates in the construction and analysis of tests; thus, for example, the ten tests most widely used by Spanish psychologists as mentioned above were all, without exception, developed within the classical framework. This fact alone makes patently clear the need for professionals to have a full understanding of the classical approach, its possibilities and its limitations.

Before looking at the logic of classical theory, it should be pointed out that its roots lie in the pioneering work of Spearman in the early 20th century (Spearman, 1904, 1907, 1913). Its one hundred or so years in the field, then, have more than earned it the right to be called classical. Spearman's early initiative is followed by rapid development, so that by 1950 the essential work is done, and it is Gulliksen (1950) who carries out the canonical synthesis of the approach. Later on it would be Lord and Novick (1968) who reformulate classical theory and open the way for the new IRT approach that we shall look at presently. But let us first consider the essence of the classical approach.

CLASSICAL LINEAR MODEL

In my own experience, after more than thirty years explaining these things to psychology students, what they find most difficult to understand is why we need a model or theory to analyze test scores. "I mean, what's the problem?", they ask themselves: "There's the test, there are the scores people have obtained in the test, some high, some low, others in the middle, so off we go, let's assign a score to each testee". But things are not so simple. Psychologists, like professionals from any other field, have to ensure that the instruments they use are measuring accurately, with little error. And this applies to any measurement instrument, be it police equipment for measuring a vehicle's speed, a tape-measure for measuring distances, or the petrol pump for measuring the number of litres delivered. All such instruments must conform to certain standards, and require some kind of indicator of the degree of precision with which they measure – and this is especially so in the case of tests, on the basis of which very important decisions affecting people's lives are made. It is not difficult to agree with this, but the problem is that when psychologists apply a test to a person, or to several people, what they obtain are the empirical or *observed* scores of the people taking the test, but this tells us nothing about the degree of accuracy of these scores; we do not know whether these observed scores are equivalent or not to the scores truly corresponding to that person in the test. It may well occur that the scores are, for example, somewhat reduced because on that particular day the person was not feeling at his or her best, or because the physical conditions in which the test was applied were not the most appropriate, or because the relations between those who applied the test and the testees left much to be desired. Psychologists, like those who construct petrol pump gauges, are obliged to guarantee that the scores on our tests are accurate, with little error. The problem is that we cannot know this by simply looking at the scores the persons obtain in the test: these scores looked at directly tell us nothing about their degree of accuracy. And this is why we have to consider them from different angles; why we have to propose some models that underlie the scores, which enable us to estimate their degree of precision. The error is mixed with the true score, like salt in seawater, or dust in straw, and to separate them we need to carry out certain processes. This is where statistical theories or models come in. There have been many models for this, but one of the most effective and parsimonious has been the classical linear model originally proposed by Spearman. Understanding the logic and functioning of the model is very simple; what is somewhat more tiresome, though not difficult, is to develop the formal aspects and deductions of the model, which constitutes the central body of psychometrics, but for this we have psychometricians. Somebody has to do it.

What did Spearman propose at the beginning of the 20th century that has been so successful in the history of Psychology? Spearman proposed a very simple, commonsense model for people's scores on tests, and which has come to be called the Classical Linear Model. It consists in assuming that the score a person obtains in a test, which we call their observed score (usually denoted by the letter X), is made up of two components: on the one hand, the person's true score on that test (T), whatever it may be, and on the other, an error (e), which can be due to many factors of which we are unaware and which we cannot control. This can be expressed formally as: $X = T + e$.

However, while this might be easily understood, we can say with some justification that we have made little progress, since if a person obtains 70 points of observed score, the model does not tell us either what their true score is or the error contained in that score. What we have, indeed, is precisely a single piece of information, the observed score (X), and two unknowns, the true score (T) and the error (e). From this point of view we have not advanced at all. We do have a score model that appears sensible and plausible, but nothing more – and nothing less: having a plausible model is the most we can ask for as a starting point. The error made on measuring some variable with a test (e) may be due to many factors, which can be found in the testee him/herself, in the context, or in the test. A quite exhaustive classification of possible error sources is provided in Stanley (1971). In order to move on, Spearman adds three assumptions to the model and one definition. Let us look at these.

The first assumption involves defining the true score (T) as the mathematical expectation of the observed score, which in formal terms can be expressed as: $T = E(X)$. What this means conceptually is that the true score of a person in a test is defined as that score they would obtain as a mean if they were to take the test an infinite number of times. The definition is a theoretical one – obviously, nobody will be asked to take a test an infinite number of times, but it seems plausible that if this were the case, that person's mean score in the test would be their true score.

In the second assumption Spearman assumes there is no relationship between the value of a person's true scores and the size of the errors that affect those scores. In other words, that the value of a person's true score has nothing to do with the error affecting that score, so that there may be high true scores with low errors or with high errors: there is no connection between the size of the true score and the size of the errors. Once more the assumption is reasonable in principle, and it can be expressed formally as: $r(v,e) = 0$.

According to the third assumption, a person's measurement errors in a test are not related to the measurement errors in a different test. That is, there is no reason to think that the errors made on one occasion will covary systematically with those made on other occasion. Formally, this assumption can be expressed as: $r(e_j, e_k)=0$.

But while these assumptions seem perfectly reasonable, they cannot be confirmed empirically in a direct fashion; it will be the deductions made on the basis of them that permit us to confirm or reject them. A hundred years since their formulation and with many empirical results behind them, we can certainly say today that Spearman's ideas have been of great utility to psychology.

In addition to the model and these three assumptions, a definition is formulated of Parallel Tests, these being understood as those tests that measure exactly the same thing but with different items. People's true scores in parallel tests would be the same, as would the variances of the measurement errors.

The linear model, then, together with its three assumptions and the definition of parallel tests proposed, constitute the central core of Classical Test Theory. A systematic course in Psychometrics consists in carrying out the corresponding deductions to reach, based on these ingredients, the formulas that permit us to estimate the degree of error in test scores, usually referred to as Test Reliability (see the work in this same issue by Prieto and Delgado, 2010). Other popular psychometric formulas can also be obtained, such as that of Spearman-Brown, which permits us to estimate the reliability of a test when it is lengthened or shortened; or the attenuation formulas that permit us to estimate the validity coefficient of a test on attenuating the measurement errors of either the test or the criterion. Nor should we overlook the formula that permits the estimation of the changes in the reliability of a test on varying the variability of the sample in which it is calculated. In sum, the classical linear model presented here, together with the assumptions and the definition of parallel tests, form the basis of all the classical formulas customarily used by psychologists who rely on tests in their professional practice. One might argue that in order to use these formulas it is not necessary to know where they come from, or understand their basis, but such an attitude would be unworthy of those psychologists who respect themselves, their science and their profession.

So, when psychologists talk about reliability and validity coefficients to indicate to their clients or users in general that the tests they are employing are accurate, with little measurement error, they should be aware that this estimation of reliability can be made thanks to this simple model and to the assumptions formulated more than a century ago.

GENERALIZABILITY THEORY AND CRITERION-REFERENCED TESTS

This classical approach has generated a range of variants, depending above all on the way the measurement error is treated. There have been numerous attempts to estimate the different components of the error, trying to break it down into its parts. Of all of these attempts, the most well known and systematic is Generalizability Theory (GT), proposed by Cronbach and cols. (Cronbach, Gleser, Nanda & Rajaratnam, 1972). It is a complex model to use, employing variance analysis for the majority of its calculations and estimations.

Another psychometric development that has emerged within the classical framework is that of Criterion-Referenced Tests (CRT). These are tests employed mainly in educational and work-related settings. Their objective is to determine whether people have mastered a specific criterion or field; thus, they set out not to discriminate between people, like the majority of psychological tests, but rather to evaluate the extent of a person's mastery in a field of knowledge called criterion, hence their name. These tests were developed from the proposal by Glaser (1963), and have had considerable influence, especially in the educational context. The classical psychometric indicators developed from the classical linear model did not adapt well to the philosophy of the construction of these new tests, and this resulted in the development of a whole set of specific psychometric techniques for calculating reliability and validity, as well as for setting the cut-off points that determine whether or not a person has mastery of the criterion evaluated (Berk, 1984; Cizek, 2001; Educational Measurement, 1994; Muñiz, 2000).

LIMITATIONS OF THE CLASSICAL APPROACH

Of the classical theory approach it can certainly be said that it is in very good health, with few doubts about its utility and efficacy; suffice to say, for example, that the vast majority of tests published in Spain, indeed practically all of them, are developed and analyzed within this framework. But if this is the case, then the obvious question is: why do we need other test theories? Or put another way: which measurement problems were not well resolved within the classical framework, giving rise to the proposal of new theories? There were in fact two basic questions not properly solved in classical theory, and which meant that psychological measurement fell short of the standard found in other empirical sciences.

Let us consider the first of these questions: within the classical framework, measurements are not invariant with respect to the instrument used. The reader may well ask what exactly this somewhat cryptic statement actually means. The answer is quite simple: if a psychologist evaluates the intelligence of three different people with a different test for each person, the results are not comparable; strictly speaking we cannot say which person is the most intelligent. This is so because the results of the three tests are not on the same scale, since each test has its own. This may surprise psychologists who habitually use the classical theory, accustomed in practice to comparing the intelligence of people that have been evaluated with different intelligence tests. In doing so, they transform the raw test scores into others calibrated, for example, in percentiles, after which they consider it appropriate to compare them. This classical procedure for solving the problem of invariance is not necessarily incorrect. However, quite apart from its scientific inelegance, it rests on quite shaky foundations, in the sense that it assumes that the normative groups in which the calibration of the different tests takes place are equivalent, and this is difficult to guarantee in practice, but without it the comparison collapses. Undoubtedly, the most desirable situation from a scientific point of view would be that the results obtained on using different instruments were on the same scale, resolving everything at a stroke. And however strange and counterintuitive it may seem, this is precisely what IRT achieves. The IRT approach represents a great advance for psychological measurement through the new psychometric concepts and tools developed within it.

The second main unresolved question from the classical framework concerns the absence of invariance of the properties of tests with regard to the persons used for estimating them. In other words, important psychometric properties of tests, such as item difficulty or test reliability, depended on the type of people used for calculating them, and this is inadmissible if measurement is to be considered rigorous. For example, item difficulty or reliability coefficients depended to a large extent on the type of sample used for calculating them. This problem would also find a suitable solution within the IRT framework.

Apart from these two broad issues there were other, minor ones of a more technical nature for which the classical theory failed to offer a satisfactory solution. For example, when a reliability coefficient for a test is offered in the classical framework, such as Cronbach's (1951) alpha coefficient, it is presupposed that the test measures all the people evaluated in the test with a given reliability, when we have more than sufficient empirical evidence that tests do not measure everyone with the same accuracy, since this accuracy depends to a large extent on the person's level in the measured variable. The new IRT framework would solve this problem by offering the Information Function, which permits the estimation of the test's reliability as a function of the person's level in the measured variable.

Apart from these central questions, IRT would generate a whole new psychometric technology that would change the *modus operandi* of psychometrics forever; see, for example, in this same special issue, the work by Olea, Abad and Barrada (2010). Nevertheless, we should make quite clear that these new models in no way invalidate the classical approach, even though they undoubtedly constitute an excellent complement which in certain circumstances provides solutions to problems not adequately covered in the classical framework. The two technologies sit perfectly side by side in the construction and analysis of tests, like cars and planes in transport: some are suitable for use in certain situations and others in other sets of circumstances.

Let us now look at the fundamental concepts on which IRT models are based.

ITEM RESPONSE THEORY (IRT)

As we pointed out in the previous section, IRT would resolve some of the serious problems of psychological measurement that could not be resolved satisfactorily within the classical framework. The price to pay for this advance was the formulation of more complex and less intuitive models than those of the classical approach, even if these did not involve a particularly high degree of difficulty. But before outlining the basics of these models, we shall provide a brief overview of their historical origins, so that the reader can place them within the history of psychology. Those who are interested in a more detailed history can consult, for example, Muñiz and Hambleton's (1992), "Half a century of item response theory".

HISTORICAL OUTLINE

In science few advances emerge suddenly, from one day to the next, without incubation. What usually happens is that there is a gradual process which at some point congeals into a recognizably new line of work. And this is more or less what happened in the case of IRT, whose origins can be traced to the pioneering work of Thurstone in the 1920s (Thurstone, 1925), which were built upon in the 1940s with the contributions of authors such as Lawley (1943, 1944) or Tucker (1946). Clearly, even in these years of total dominance of the Classical Theory, the new perspective that would come to be known as IRT was taking its early steps. These are the remote origins, but it would be the notable psychometrician Frederic Lord (1952) who, in his doctoral thesis supervised by Gulliksen (the great synthesist of Classical Theory), laid the true foundations of IRT. Birnbaum, in the 1950s, made some important contributions, before the Danish mathematician Rasch (1960) proposed his now famous one-parameter logistic model. We might well consider this date as the point at which IRT took off, but it should not be overlooked that at this time we are still in merely theoretical and statistical territory, a long way short of the practical application of these new models. The great boost in this regard would come from Lord and Novick (1968) in their famous work, in which they devoted five chapters to the topic. In the wake of their book, research on IRT models began to predominate in psychometrics, and continues to do so today. It was not long after Lord and Novick's book that there began to appear the computer programs necessary for using IRT models, such as BICAL and LOGIST in 1976, BILOG in 1984, MULTILOG in 1983, and many others. In 1980 Lord would publish another influential book (Lord, 1980) on the applications of IRT. From then until now enormous progress has been made, and we can safely say that IRT now dominates the scene in psychometrics. An introduction to IRT in Spanish can be found, for example, in Muñiz (1997a); a recommended work in English is the book by Hambleton, Swaminathan and Rogers (1991). Let us now consider the assumptions and models of IRT.

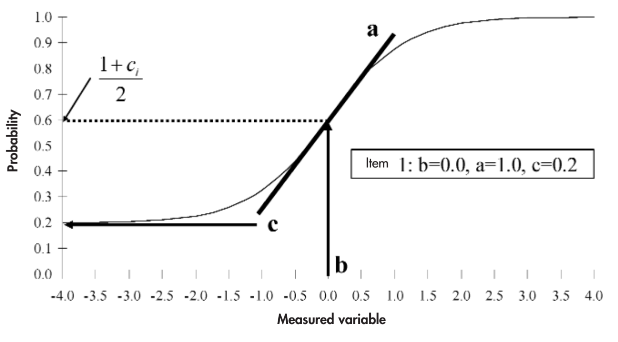
ASSUMPTIONS

To resolve the problems mentioned above for which the classical framework did not find a good solution, IRT would have to make some stronger and more restrictive assumptions than those of Classical Theory. The key assumption in IRT models is that there is a functional relationship between the values of the variable measured by the items and the probability of the item being answered correctly, this function being referred to as the Item Characteristic Curve (ICC) (Muñiz, 1997a). An example of this can be seen in Graph 1. Note that as the values of the measured variable, called θ, increase, so does the probability of the item being answered correctly, P(θ). The values of the measured variable, whatever it may be, range from minus infinity to plus infinity, whilst in classical theory the values depended on the scales of each test, ranging from the minimum to the maximum value obtainable on the test.

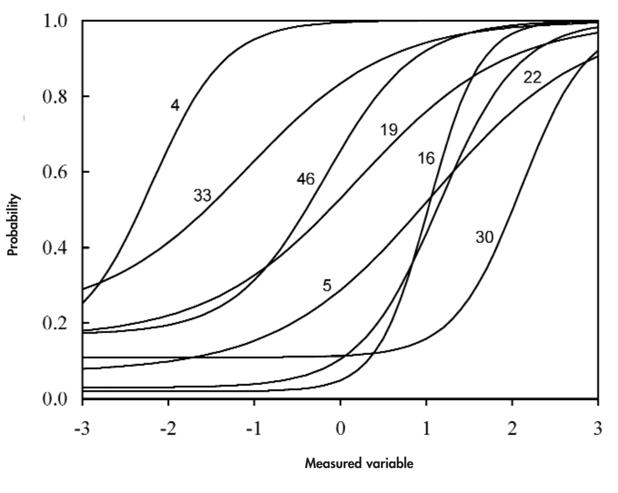
The specific form of the ICC is determined by the values of three parameters: *a*, *b* and *c*. Parameter *a* is the discrimination index of the item, *b* is the item difficulty and *c* is the probability of the item being answered correctly by chance. As the parameters take different values, the form of the curve changes, as can be seen in Graph 2.

Naturally, the parameter values are calculated on the basis of the data obtained on applying the items to a large and representative sample of individuals. Sophisticated computer programs are necessary for these calculations, so that it is not surprising that IRT models were not widely used until the advent of powerful computers.

GRAPH 1
ITEM CHARACTERISTIC CURVE (ICC)



GRAPH 2
CHARACTERISTIC CURVES OF EIGHT ITEMS



The majority of IRT models, and certainly the most popular of them, assume that the items constitute a single dimension, so that before using these models it must be ensured that the data fulfil this condition – that they are one-dimensional. This represents a considerable restriction for their use, as it is well known that much of the data handled by psychologists is not essentially one-dimensional, even though it is true that the models still work quite well when the data are not strictly one-dimensional, that is, they are reasonably robust to moderate violations of one-dimensionality (Cuesta & Muñiz, 1999).

A third assumption of IRT models is so-called Local Independence, which means that to use these models the items must be independent of one another, that is, the response to one of them cannot be conditional upon the response given to other items. In reality, if one-dimensionality is fulfilled, so is Local Independence, so that the two assumptions are sometimes treated jointly.

MODELS

With the assumptions indicated, according to whether we choose for the Characteristic Curve of the items one mathematical function or another, we shall have different models, so that we tend to talk about IRT models, in the plural. Theoretically, there would be an infinite number of possible models, since there are plenty of mathematical functions to choose from, but the most widely used functions, for various reasons, are the logistic function and the normal curve. The logistic function has many advantages over the normal curve, since it gives similar results and is much easier to handle mathematically, so that the three IRT models most commonly used are the logistic models, which adopt the logistic function as the Characteristic Curve of the items. If we take into account only the item difficulty (parameter *b*), we have the one-parameter logistic model, or Rasch model, after the author who proposed it (Rasch, 1960). If in addition to the difficulty we take into account the discrimination index of the items (parameter *a*), we are looking at the two-parameter logistic model, and if we also add the probability of getting the item right by chance (parameter *c*), we have the three-parameter logistic model. This is the most general of the three; indeed, the other two are particular cases, so that when parameter *c* is zero we have the two-parameter model, and when, moreover, parameter *a* is equal for all the items, it becomes converted into the Rasch model. Below is the formula of the three-parameter logistic model, where P(θ) is the probability of getting the item right, θ is the score on the measured variable, *a*, *b* and *c* are the three parameters described, *e* is the base of the naperian logarithms (2,72) and D is a constant with a value of 1.7.

$$P(\theta) = c + (1-c)\ [e\ Da(\theta-b)\ /(1+e\ Da(\theta-b)\ )]$$

Currently there are more than 100 IRT models, which are used according to the type of data being handled. Thus, we have models for Likert-type scales, for dichotomous data, or for multidimensional data. A comprehensive classification and review of the models can be found in the book by Van der Linden and Hambleton (1997).

TABLE 1
DIFFERENCES BETWEEN CLASSICAL THEORY AND ITEM RESPONSE THEORY

Aspects	Classical Theory	Item response Theory
Model	Linear	Non-linear
Assumptions	Weak (easy to fulfil with the data)	Strong (difficult to fulfil with the data)
Measurement invariance	No	Yes
Invariance of test properties	No	Yes
Score scale	Between zero and the maximum test score	Between - ∞ and ∞
Emphasis	Test	Item
Item-test relationship	Not specified	Item Characteristic Curve
Item description	Difficulty and Discrimination Indices	Parameters a, b, c
Measurement errors	Standard error of measurement common to whole sample	Information function (varies according to aptitude level)
Sample size	Can work well with samples of between 200 and 500 participants approx.	More than 500 participants recommended, but depends on model.

COMPARISON OF CLASSICAL THEORY WITH IRT

Table 1, taken from Muñiz (1997a), summarizes the differences and similarities between the classical approach and IRT.

BY WAY OF CONCLUSION

The aim of this article was to present in a non-technical manner to professional psychologists – readers of *Papeles del Psicólogo* – the most influential theories in the construction and analysis of tests: Classical Test Theory and Item Response Theory. It is my hope that this overview of the basics will help readers to better understand and interpret the psychometric data customarily provided in relation to tests. It is also to be hoped that it would encourage them to refresh their psychometric knowledge and to consider in more depth some new aspects relevant to their professional practice. Everything related to psychological measurement has evolved extremely rapidly in recent decades, resulting in significant advances, and psychologists must keep abreast of these developments if they are to avoid falling behind in the area of psychological assessment. Without precise and rigorous assessment we cannot make accurate diagnoses, which in turn are essential for any kind of effective intervention.

REFERENCES

Anastasi, A. & Urbina, S. (1998). *Los tests psicológicos* [*Psychological tests*]. Mexico: Prentice Hall.

Andrés-Pueyo, A. (1996). *Manual de psicología diferencial* [*Handbook of differential psychology*]. Madrid: McGraw Hill.

Berk, R. A. (Ed.) (1984). *A guide to criterion referenced test construction*. Baltimore, MD: The Johns Hopkins University Press.

Binet, A. & Simon, T. H. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux [New methods for diagnosing the intellectual level of the abnormal]. *L'année Psychologique, 11*, 191-244.

Carretero-Dios, H., & Pérez, C. (2005). Normas para el desarrollo y revisión de estudios instrumentales [Standards for the development and review of instrumental studies]. *International Journal of Clinical and Health Psychology, 5*, 521-551.

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

Cizek, G. J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives*. London: LEA.

Colom, B. R. (1995). *Tests, inteligencia y personalidad* [*Tests, intelligence and personality*]. Madrid: Pirámide.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

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

Cuesta, M. & Muñiz, J. (1999). Robustness of item response logistic models to violations of the unidimensionality assumption. *Psicothema*, Vol. 11, 175-182

Downing, S. M. & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Educational Measurement: Issues and Practice (1994). Special issue on thirty years of criterion-referenced tests. Vol. 13, nº 4.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519-521.

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

Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*. Beverly Hills, CA: Sage.

Juan-Espinosa, M. (1997). *Geografía de la inteligencia humana* [*The Geography of Human Intelligence*]. Madrid: Pirámide.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61*, 273-287.

Lawley, D. N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh, 62*, 74-82.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs, nº 7.*

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: LEA.

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

Morales, P., Urosa, B., & Blanco, A. B. (2003). *Construcción de escalas de actitudes tipo Likert* [*The construction of Likert-type attitudes scales*]. Madrid: La Muralla.

Muñiz, J. (1997a) *Introducción a la teoría de respuesta a los ítems* [*An introduction to item response theory*]. Madrid: Pirámide.

Muñiz, J. (1997b). Aspectos éticos y deontológicos de la evaluación psicológica [Ethical and deontological aspects of psychological assessment]. In A. Cordero (ed.), *La evaluación psicológica en el año 2000* [*Psychological assessment in the year 2000*]. Madrid: Tea Ediciones.

Muñiz, J. (2000). *Teoría Clásica de los Tests* [*Classical Test Theory*]. Madrid: Pirámide.

Muñiz, J. (2005). Classical test models. In B. S. Everitt and D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. Chichester: John Wiley and Sons. (Vol. 1, pp. 278-282).

Muñiz, J. & Bartram, D. (2007). Improving international tests and testing. *European Psychologist, 12*, 206-219.

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J. R., & Zaal, J. (2001). Testing practices in European countries. *European Journal of Psychological Assessment, 17(3),* 201-211.

Muñiz, J. & Fonseca-Pedrero, E. (2008). Construcción de instrumentos de medida para la evaluación universitaria [The construction of measurement instruments for university assessment]. *Revista de Investigación en Educación, 5*, 13-25.

Muñiz, J. & Hambleton, R. K. (1992). Medio siglo de teoría de respuesta a los ítems [Half a century of item response theory]. *Anuario de Psicología, 52(1)*, 41-66.

Olea, J., Abad, F.J., & Barrada, J.R. (2010). Tests informatizados y otros nuevos tipos de tests [Computerized tests and other new types of test]. *Papeles del Psicólogo, 31(1)*, 97-107

Pereña, J. (2007). *Una tea en la psicometría española* [*A torch in Spanish psychometrics*]. Madrid: Tea Ediciones.

Prieto, G. & Delgado, A. (2010). Fiabilidad y validez [Reliability and validity]. *Papeles del Psicólogo, 31(1), 67-74*.

Prieto, G. & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España [A model for assessing the quality of tests used in Spain]. *Papeles del Psicólogo, 77*, 65-71.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute for Educational Research.

Schmeiser, C. B. & Welch, C. (2006). Test development. In R. L. Brennan (Ed.), *Educational Measurement (4th ed.)* (pp. 307-353). Westport, CT: American Council on Education/Praeger.

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

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

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

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational Measurement*. Washington: American Council on Education.

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

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

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, nº 1.

Thurstone, L. L. & Thurstone. T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs*, nº 2.

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-13.

Van der Linden, W. J. & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wissler, C. (1901). Correlation of mental and physical traits. *Psychological Monographs, 3*, nº 16.