Psychometrics

PSICOMETRIA

PSICOMETRÍA

Luiz Pasquali[1]

ABSTRACT
Psychometrics has foundations on the theory of measurement in Sciences and is aimed at explaining the meaning of responses provided by subjects submitted to a series of tasks, and proposing techniques for the measurement of mental processes. This article presents concepts and models of modern psychometrics and discusses the validity and reliability parameters of the applied tests.

KEY WORDS
Psychometrics.
Reproducibility of results.
Validity of tests.
Validation studies.

RESUMO
A psicometria fundamenta-se na teoria da medida em ciências para explicar o sentido que têm as respostas dadas pelos sujeitos a uma série de tarefas e propor técnicas de medida dos processos mentais. Neste artigo são apresentados os conceitos e modelos da psicometria moderna e discutidos os parâmetros de validade e precisão dos testes.

DESCRITORES
Psicometria.
Reprodutibilidade dos testes.
Validade dos testes.
Estudos de validação.

RESUMEN
La Psicometría se fundamenta en la teoría de la medida en las ciencias buscando explicar el sentido en las respuestas de los que fueron sujetos a una serie de tareas, además de proponerse técnicas de medida de sus procesos mentales. En este artículo son presentados los conceptos y modelos de psicometría moderna, así como son discutidos los parámetros de validez y precisión de los testes.

DESCRIPTORES
Psicometría.
Reproducibilidad de resultados.
Validez de las pruebas.
Estudios de validación.

[1] Researcher Professor Associated with the University of Brasilia. Brasilia, DF, Brazil. luiz.pasquali@pq.cnpq.br

Received: 15/06/2008
Approved: 15/12/2008

INTRODUCTION

Measurement in psychosocial sciences

Psychometrics is etymologically represented as the theory and technique of measuring mental processes, and is especially applied in the fields of psychology and education. It is grounded in the general theory of measurement in sciences, or else, in the quantitative method whose major characteristic is the fact that it represents the knowledge of nature in a more precise way in comparison with the application of common language to describe the observation of natural phenomena.

Psychometrics historically stems from the psychophysics of the Germans Ernst Heinrich Weber and Gustav Fechner. The British Francis Galton also contributed to the development of psychometrics by creating tests to measure mental processes; by the way, he is considered as the creator of psychometrics. However, it was the inventor of the multiple factorial analyses, Leon Louis Thurstone, who enlivened psychometrics, making it different from psychophysics. Psychophysics was defined as the measurement of directly observed processes, or in other words, the organism's stimulus and response, while psychometrics consists in measuring the organism's behavior by means of mental processes (law of comparative judgment).

Measurement in sciences has raised diatribes among researchers, particularly in the field of social sciences. Nonetheless, the most accepted definition among researchers was given by Stanley Smith Stevens in 1946. He used to say that to measure meant *to assign numbers to objects and events in accordance with given rules*[1]. The assignment rules to such numbers are defined by the proposal of the same author concerning the four measurement levels or measurement typologies, which are: nominal, ordinal, interval, and ratio.

The nominal measurement is the one that applies numbers to nature phenomena, keeping exclusively the axioms of number identity, that is, the number is employed only as a numeric or graphic symbol. When applying the number, the ordinal typology *saves* the axioms of order, that is to say, the major characteristics of the number, or its magnitude (by definition, a given number is greater or smaller than, not only different from or better than the other exactly because its value is intrinsically higher or lower than any other). The other typologies point to axioms of additionality. The axiom history was detailed by Whitehead and Russell between 1910 and 1913, and again in 1965, in their book Principia Mathematica, where they describe the 27 famous axioms of the mathematical number[2].

PSYCHOMETRICS: CONCEPT AND MODELS

Modern psychometrics can be traced back to two sources: the classical test theory (CTT), and the item response theory (IRT). CTT has been axiomatized by Gulliksen[3] and IRT was initially elaborated by Lord[4] and Rasch[5], and finally axiomatized by Bimbaum[6] and Lord[7].

In a general sense, psychometrics attempts to explain the meaning of responses given by subjects in a series of tasks typically named as *items*. The CTT is aimed at explaining the total final result, that is, the sum of responses provided to a series of items, expressed by the so-called total score (S). For instance, the S in a test of 30 capability items would be the sum of correctly responded items. If the value of 1 were given to each correct item and 0 to each incorrect one, and the subject reached 20 correctly and 10 incorrectly responded items, this person's score S would be 20. The CTT, then, asks itself: what does this total 20 mean to the subject? The IRT, on the other hand, is not interested in the test total score; it is specifically aimed at each one of the 30 items and wants to know what the probability is and what the factors that influence this probability are regarding every individual item's correctness and incorrectness (in capability tests) or acceptance or rejection (in preference tests: personality, interests, attitudes). In such a way, the CTT is interested in producing quality tests, while the IRT is focused on developing quality tasks (items). At the end, therefore, we have either valid tests (CTT) or valid items (IRT), and those results will build as many valid tests as desired, or the amount of tests allowed by the items. Thus, the richness of the psychological or educational assessment within the IRT's scope of action consists in building store rooms of valid items that evaluate latent traits - these store rooms are called *item bank*, aimed at elaborating countless numbers of tests.

> Psychometrics attempts to explain the meaning of responses given by subjects in a series of tasks typically named as *items*.

The CTT model was elaborated by Spearman and detailed by Gulliksen, as follows:

$$T = TS + E$$

where,

T = subject's total or empirical score, which is the sum of all items achieved by the test;

TS = true score, which is the real magnitude of what the test wants to measure in the subject; that score will be the S itself, in case there is no measurement error;

E = the error of the measurement.

In this way, the empirical score is the sum of the true score and the error; consequently, E = T – TS, and TS = T – E.

Figure 1 shows the relationship among these various elements of the empirical score, where the union between the true (TS) and the error (ES) score can be observed; that is to say, the subject's empirical or gross score (T – test result known as the Tau score - τ) is comprised of two components: the subject's real or true score (TS) in what the test intends to measure, and the error score (ES) of the measurement, which is always present in any empirical operation. In other words, we are assuming here that as the subject's gross score differs from his true score, it is the error that accounts for such a disparity; this difference, then, is the error's concept itself.

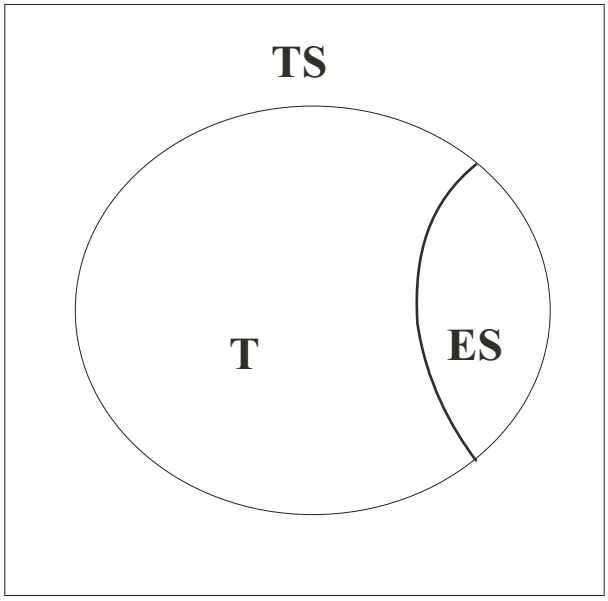

Figure 1 - The true score (TS) components

Hence, the CTT's ultimate challenge is to elaborate strategies (statistical ones) to either control or evaluate E's magnitude. Errors are provoked by a wide range of alien factors identified by Campbell and Stanley[8], such as the test's own deficiencies, stereotypes and biases of the subject, historical factors, and random historical and environmental factors.

On the other hand, the IRT model works with latent traits and adopts two fundamental axioms:

The subject's performance in a task (test item) is explained by a set of factors or latent traits (capabilities, skills, etc.). The performance is the effect; latent traits are the cause.

The relationship between the performance in a task and the set of latent traits can be described by a crescent monotonic equation called ICC (Item Characteristic Function or Item Characteristic Curve). It is exemplified in Figure 2, which shows that subjects with higher capability will most probably respond correctly to the item and vice-versa (θ_i is the capability and $P_i(\theta)$ the correct response probability given to the item).

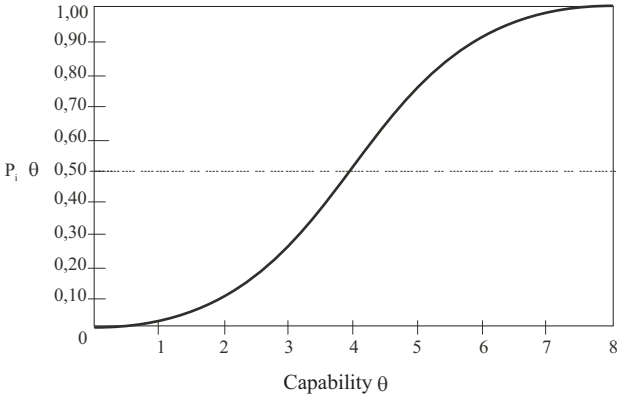

Figure 2 - The item's characteristic curve

The IRT is concretely affirming the following: the subject is given a stimulus or a series of stimuli (such as, items of a test) and he/she responds to it/them. From the responses provided by the subject, that is, taking into account the analysis of his/her responses to the specified items, we can deduce on the subject's latent trait, hypothesizing relationships between the subject's observed responses and the level of his/her latent trait. These relationships can be expressed by means of a mathematical equation that describes the type of function taken on by these relationships.

In fact, only a limited number of mathematical models are able to express such relationships, depending on the type of applied mathematical function and/or the number of parameters that one wants to find out for the item. A remarkable advantage IRT has over the classical theory concerning the models it uses is that the models employed by the IRT allow for disconfirmation. In effect, the demonstration of compatibility between the model and the data (model-data goodness-of-fit) is a necessary step towards this theory's procedures. Specialized statistical packages are made necessary in order to perform the IRT, as they are abundant in the market[a].

TEST PARAMETERS: VALIDITY AND RELIABILITY

The two most important parameters of measurement or test legitimacy used both by the CTT and the IRT are the validity and reliability.

Test validity

In the context of psychosocial sciences, validity is a typically discussed measurement parameter. It is not a typical issue in physical sciences, although the parameter would be applicable in certain physical situations. The physical sciences' major concern is centered on the issue of reliabil-

(a) Two of the most used packages are the BILOG for capability tests, and the PARSCALE for personality tests.

ity, or the so-called instrument calibration. This measurement issue is also relevant to psychosocial sciences, although it conceptually has nothing to do with the validity issue.

This is because validity refers to the congruence between the instrument being used for measurement and the property under evaluation and not regarding the accuracy that describes the object's property. In physics, the instrument is a physical object that measures physical properties; then, it seems easy to acknowledge whether or not the object's measuring property is congruous with the measured object's property. Take the object's *length* property, for example. The instrument that measures this property (length), the meter, applies its length property in order to measure another object's length; so, we are not matching length with length as univocal terms. There is no need to prove that the meter's *length* property is congruous with the same property in the measured object; terms are univocal, conceptually equivalent, and identical.

It is less clear, however, when the astronomer measures the galactic *speed* property of approximation or withdrawal via Doppler Effect, where approximation/withdrawal of the galaxy's light spectral lines would be the measurement instrument. Here, we actually have a problem to validate the measurement instrument; the question is: is it or is it not true that spectral line distances have to do with the speed of galaxies? Such an inference can be made, but it has to somehow be empirically demonstrated, that is, at least its consequences should be indicated, as well as all the derived, derivable, or verifiable hypotheses. In this specific case, the problem of measurement precision is related to the preciseness of the distance measurements of the oscilloscope's spectral lines, whereas the validity is related to whether or not the measurement of spectral line distances, regardless its accuracy and perfection, has something to do with the galaxy's withdrawal speed. In other words, the validity in such case refers to the demonstration of compatibility (legitimacy) in the representation or modeling of galactic speed via spectral line distances.

Validity refers to the congruence between the instrument being used for measurement and the property under evaluation and not regarding the accuracy that describes the object's property.

This astronomy case illustrates what typically occurs with psychosocial sciences measurements, and consequently turns the evidence of instrument validity in these sciences into an essential and crucial aspect; to show the validity of instruments in these sciences is a sine qua non condition. This is particularly the case of the above-mentioned focuses that deal with the psychological concept of *latent trait*, where the correspondence (congruence) between latent trait and its physical representation (behavior) must be demonstrated. It is not incidental, therefore, that the problem of validity has taken a central role in the measurement theory in the history of psychology; in fact, it is its basic and indispensable parameter.

Psychometrics manuals usually define the validity of any given test by certifying whether or not the test measures what it is supposed to measure. Although this definition may sound like a tautology, when the psychometric theory that admits the latent trait is taken into account it proves to be not. This definition clearly states that whenever behaviors (items) are measured - and behaviors are the physical representation of the latent trait - the latent trait itself is being measured. This supposition is only possible when an existing previous trait theory supports the behavioral representation as a deductible hypothesis for the theory. The test validity (the hypothesis), therefore, will be established by the empirical testing of the hypothesis verification. At any rate, this is the scientific methodology. Hence, the current psychometrics practice of intuitively grouping a series of items and statistically verifying a posteriori what they are measuring becomes quite unusual. The emphasis in the formulation of the trait theory used to be quite weak in the past; under the influence of the cognitive psychology, psychometrics is fortunately retaking this emphasis, bringing it back to its relevant place.

The classical psychometrics, by the way, understands *what supposedly has to be measured* as the *criterion*, which is represented by a parallel test. Thus, the *what* is the latent trait in the cognitivistic conception of psychometrics, and it is the criterion (score in the parallel test) in the behavioralist perspective.

The validation process of any given test

> begins with the formulation of detailed definitions of specific traits or constructs, derived from psychological theory, previous research, or systematic observation and analysis of the relevant domains of behavior. The items of the test, then, are prepared in order to fit the construct's definitions. Next, empirical analysis of the items are implemented, and the more efficient (i.e., valid) items are finally selected from the initial sample of items[9].

Although it constitutes the core point of psychometrics, the validation of the trait's behavioral representation, or the test's representation, brings about significant difficulties that are located in three levels in the process of elaborating the instrument, namely: the theory level, the information empirical collection level, and the statistical analysis of information properly said.

The most significant difficulties are probably centered at the level of theory. As a matter of fact, the psychological theory is still found in an embryonic state, and so it virtually lacks any level of axyomatization. As a result, a wide scope of theories arises, even contradictory ones. It is worth remembering that we have several theories, such as behaviorism, psychoanalysis, existentialist psychology, dialectical psychology, and others; when existing simultaneously,

they postulate irreducible principles among the various theories; they also can weakly combine principles within the same theory, or even present an insufficient aspect that is unable to develop useful hypothesis for the psychological knowledge. This confused perspective takes place in the theoretical field of the constructs, that is, in the formulation of clear and accurate hypothesis to either test or postulate useful psychological hypothesis. Even when there is success in the operationalization process, the empirical data collection will not be exempt of difficulties, such as, for example, the unequivocal definition of criteria groups where these constructs can be ideally studied. Problems are found even at the level of the statistical analysis. According to the elaboration logic of the instrument, the hypothetical verification of the construct's representation legitimacy is performed by means of analyses such as the factorial analysis (confirmatory), which attempts to identify the previously operationalized constructs of the instrument in the empirical data. But, the factorial analysis happens to make some strong postulations that not always match the reality of facts. For instance, the factorial analysis assumes that subjects' responses to the instrument's items are determined by a linear relationship these subjects have with the latent traits. The rotation of axles is another serious problem, allowing for countless numbers of factors related to the same instrument[10].

Having these difficulties in mind, psychometricists call upon a series of techniques in order to make possible the demonstration of the instrument's validation. These techniques can essentially be reduced to three large classes (the trinitarian model): construct validation; content validation; and criterion validation[11,12].

The construct validation, or concept validation, is deemed as the most fundamental form of validating psychological instruments, and this is quite reasonable, since it constitutes the direct way of verifying the hypothesis of the behavioral representation legitimacy of latent traits; therefore, it is connected with the psychometrics theory defended here. Historically, the *construct* concept was inserted into psychometrics through the American Psychological Association Committee on Psychological Tests, which functioned between 1950 and 1954, and whose results later became technical recommendations for psychological tests[12].

The concept of construct validity was elaborated by the classical article by Cronbach and Meehl[13], Construct validity in psychological tests, although the concept was already part of history under other names, such as intrinsic validity, factorial validity, and face validity. These various terms show the confusing notion expressed by constructs. In spite of the fact that Cronbach and Meehl attempted to clarify the concept of construct validity, they still define them as the characteristic that any test has of measuring an attribute or quality that has not been *operationally defined*[13]. They recognize, however, that the construct validity required a new scientific focus. In fact, to define validity as they did sounds a bit uncommon to sciences, as operationally non-defined concepts are not susceptible to scientific knowledge. Concepts or constructs are scientifically researchable only when they are liable for adequate behavioral representation. Otherwise, they will only be metaphysical, non-scientific concepts. The problem stemming from the general synthetic attitude of psychometricists of then is that whenever the construct validity had to be defined, the researchers started from the test, that is, from the behavioral representation, instead of beginning with the psychometric theory grounded on the elaboration of the construct's theory (or the latent trait theory). The obstacle is not to identify the construct from any existing representation (test), but to find out whether or not the representation (test) constitutes a legitimate, adequate representation of the construct. This focus demands quite a close collaboration between psychometricists and the cognitive psychology[14]. The construct validity of any given test can be dealt with in several angles: the construct's behavioral representation analysis; the hypothetical analysis; and the IRT's information curve[15-16].

The criterion validity of a test consists of the efficiency level it has to predict the specific performance of a subject. The subject's performance thus becomes the criterion against which the measurement achieved by the test is assessed. The subject's performance must obviously be measured/assessed through techniques that are independent on the planned test itself.

There are two distinctions for a test's criterion validity: (1) predictive validity, and (2) concurrent validity. The core difference between both is basically the matter of time between the information collection of the test to be validated, and the information collection of the criterion. If both collections are performed almost simultaneously, the result will be a concurrent validity; if the data about the criterion are collected after the test's information collection, the result will be the predictive validity. The fact that the information is simultaneously reached, or reached further to the test itself, is not a technically relevant factor towards the validity of the test. The relevance is located in the determination of a valid criterion. Here the central nature of this type of test validation is situated, as follows: (1) to define an adequate criterion, and (2) to measure the criterion in a valid, independent way, regardless the test itself.

As per the criteria adjustment, we can affirm that there is a series of them that are usually employed, such as:

1) Academic performance. Perhaps this used to be, or still is the most applied criterion to validate intelligence tests. It consists in the achievement of the students' school performance by means of teachers' grades, by the students' general academic average, by the academic honors received by students, or even by the teachers' or colleagues' purely subjective assessment regarding these students' *intelligence*. Despite being broadly used, this criterion has been

similarly quite criticized mainly due to the deficiency of its assessment process. It is widely known that teachers are generally tendentious in attributing grades to students; this bias is not always a conscious act, but it stems from their attitudes and sympathies towards this or that student. Teachers could overcome this challenge quite easily if they were used to apply performance tests based on content validity, for instance. As this is quite a laborious task, teachers typically do not make efforts towards validating (content validity) the students' academic tests.

In this context, the subject's schooling level is also applied as an academic performance criterion: advanced, repeating, and dropping out subjects. Supposedly, those who keep a regular study, or those who are academically advanced proportionally to their ages have more intelligence. Evidently, not only the issue of intelligence must be worked out in this argument, but also several other social factors, personality aspects, etc., which makes this quite an ambiguous, deceitful criterion.

2) Performance in specialized training. It refers to the performance obtained in training courses under specific situations (musicians, pilots, mechanical or specialized electronic activities, etc.). At the end of this training process a typical assessment takes place, producing useful data that will serve as criteria for the students' performance. The critical observations uttered for point 1 are also replicable in this paragraph.

3) Professional performance. In this case, test outcomes are compared with the subjects' success/failure, or their quality level in the work environment. Hence, a test of mechanical ability can be implemented against the mechanical performance of subjects in a given work place. Mapping out the quality of the performance of subjects in service, again, is evidently quite a difficult task.

4) Psychiatric diagnosis. This method is quite used to validate personality/psychiatric tests. The criteria groups are comprised of the results of the psychiatric assessment that settles clinical categories: normal versus neurotic, psychopath versus depressive, etc. Again, it is very hard to adequate the psychiatrists' assessments.

5) Subjective diagnosis. Assessments performed by colleagues and friends can be a basis for the establishment of criteria groups. This technique is employed, above all, in personality tests, where more objective assessments are hardly achieved. Thus, subjects place their colleagues in categories, or score personality traits (aggressiveness, cooperation, etc), based on the experience of their living together. Needless to say that there are enormous hardships produced by these assessments in terms of objectivity; nonetheless, the application of a large number of judges can diminish the subjective biases of these evaluations.

6) Other available tests. The outcomes achieved by means of another valid test that predicts the same performance of the test to be validated can serve as a criterion to determine the validity of the new test. Here's an obvious question: what is the purpose of creating another test if an existing one validly measures what it is supposed to measure? The answer is based upon a sense of economy, that is, one makes use of a test that demands a longer length of time to be responded or assessed as a criterion to validate another test that spends a lower amount of time.

In case of this last type of validity method, two distinct situations must be met. First, whenever there are provably validated tests for the measurement of any trait, they certainly constitute a criterion against which a new test can be safely validated. Nevertheless, when tests accepted as definitely validated do not exist for the assessment of a latent trait, the application of the contending validity is extremely precarious. This situation is unfortunately the most common one. As a matter of fact, there are available tests to measure practically anything, as attested by the Buro's Mental Measurement Yearbooks, which are periodically published and contain thousands of existing psychological tests in the market. In this case, these tests can be used as validation criteria, but the risk is excessively high due to the fact that a test whose validity is minimally questionable is being employed as a criterion.

We can conclude that the concurrent validity only makes sense if provably valid tests can serve as a criterion against which one wants to validate a new test, and that this new test have some advantages over the previous one (such as, for instance, saving time, etc.).

A frustrating issue stands out at the end of this study on criterion validity processes. If the researcher has employed all his ability to build a test, under the highest degree of control possible, why would he validate this task-test against lower measures, represented by the measurement of various criteria presented here? Is it reasonable to validate supposedly superior measurements using a poorer measurement?[17]. The criticisms of both Thurstone in 1952 and above all those of Cronbach and Meehl in 1955(13,18) replaced the criterion validity of the psychological tests' validation panacea technique for the construct validity. However, these criteria can be deemed as good and useful towards the criterion validation. The significant difficulty in almost all of them is located in the demonstration of their measurement adjustment; in other words, these measurements are generally precarious, thus leaving much doubt on the test validation process. Nonetheless, there are well-known examples of validated tests through this method, such as the MMPI (Minnesota Multiphasic Personality Inventory).

A test's content validity is comprised of verifying whether or not the test constitutes a representative sample of a finite universe of behaviors (domain). It is applicable whenever a finite universe of behaviors can be delimited a priori, such as the case of performance tests that intend to cover a content that is delimited by a specific programmatic course[11].

Test reliability

The reliability or trustworthiness parameter of tests is referenced by a long and heterogeneous series of names. Some of those names stem from the own concept of this parameter; in other words, these terms attempt to express what they really represent to the test. These names are, mostly: preciseness, trustworthiness, and reliability. Other names of this parameter result more directly from the type of technique applied in the empirical collection of information, or the statistical technique employed in the analysis of the collected empirical data. Among these names we mention the following: stability, steadiness, equivalence, internal consistence.

Trustworthiness, or reliability of a test refers to the major characteristic it must display, namely, the errorless measurement; hence, we have the terms *preciseness*, *reliability*, and *trustworthiness*. An errorless measurement means that the same test that measures the same subjects in different occasions, or equivalent tests that measures the same subjects in the same occasion, produce identical outcomes; in other words, the correlation of both measurements must score 1. However, as the error is always present in any measurement, the further this correlation withdraws from 1, the bigger the measurement error will be. The reliability analysis of a psychological instrument precisely shows how much the same instrument withdraws from the ideal 1 correlation, determining a close-to-1 coefficient, so that the error probability is lower.

Tests' trustworthiness problem used to be a favorite issue for classical psychometrics, where the statistical estimation paraphernalia for this parameter grew up the most; but it lost importance within modern psychometrics in favor of the validity parameter. Anyway, within CTT, the trustworthiness coefficient, r_{tt}, is statistically defined as the correlation between the scores of the same subjects in two parallel ways of a test, T1 and T2. Hence, the trustworthiness coefficient is defined as the co-variance function [Cov(T1,T2)] between the test formats by means of their own variances ($S^2_{T_1}$ e $S^2_{T_2}$), that is, $r_{tt} = \frac{S^2_V}{S^2_T}$

where,

r_{tt} : reliability coefficient

S^2_V : test true variance

S^2_T: test total variance

There are practically two statistical techniques to decide the accuracy of a test, that is, the correlation and the analysis of the internal consistency.

The correlation technique is applied for test-retest and test parallel format conditions. Both cases show the outcomes of the same subjects that were submitted to the same test in two different occasions, or responded to two parallel formats in the same test. The reliability index, in this case, simply consists of a bi-varied correlation between both scores concerning the same subjects.

The internal consistency analysis demands a complex apparatus of statistical techniques that are finally reduced to two situations: dividing the test in shares - more commonly in two halves - with a subsequent correction made by the Spearman-Brown prediction formula, and several alpha coefficient techniques, being the Cronbach alpha the most widely known of them all. Here, only one test is applied in only one occasion; analyses consist of verifying the internal consistency of the items that compose the test. It is, therefore, an accuracy estimation, whose logic is as follows: if the items *understand* themselves, that is, covariate, in a given occasion, they will thus *understand* each other in any other occasion throughout the test.

CONCLUSION

In order to guarantee that tests will present the scientifically required quality parameters, the American Psychological Association (APA) established the Standards for Educational and Psychological Testing, with several editions since 1985.

REFERENCES

1. Stevens SS. On the Theory of Scales of Measurement. Science. 1946;103(2684):677-80.

2. Whitehead AN, Russell B. Principia mathematica. Cambridge: Cambridge University Press; 1910-1913, 1965. 3 v.

3. Gulliksen H. Theory of mental tests. New York: Wiley; 1950.

4. Lord FM. A theory of test scores. Iowa (IA): Psychometric Society; 1952. (Psychometric Monograph, n. 7).

5. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research and St. Paul; 1960.

6. Birnbaum A. Some latent trait models and their use in inferring and examinee's ability. In: Loed FM, Lord MR. Novick, statistical theories of mental test scores. Reading: Addison Wesley; 1968. p.17-20.

7. Lord FM. Applications of item response theory to practical testing problems. Hillsdale: Erlbaum; 1980.

8. Campbell DT, Stanley J. Experimental and quasi-experimental designs for research. Skokie: Rand McNally; 1973.

9. Anastasi A. Evolving concepts of test validation. Ann Rev Psychol. 1986;37(1):1-15.

10. Pasquali L, organizador. Instrumentos psicológicos: manual prático de elaboração. Brasília: LabPAM/IBAPP; 1999.

11. Pasquali L. Análise fatorial para pesquisadores. Porto Alegre: Artmed; 2005.

12. American Psychological Association (APA).Technical recommendations for psychological tests and diagnostic techniques. Washington; 1954.

13. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52(4):281-302.

14. Pasquali L.Validade dos testes psicológicos: será possível reencontrar o caminho? Psicol Teor Pesq. 2007;23(n.esp): 99-107.

15. Pasquali L. Psicometria: teoria dos testes na psicologia e na educação. Petrópolis: Vozes; 2004.

16. Pasquali L. TRI – Teoria de Resposta ao Item: teoria, procedimentos e aplicações. Brasília: LabPAM/UnB; 2007.

17. Ebel RL. Must all tests be valid? Am Psychol. 1961;16(10): 640-7.

18. Thurstone LL.The criterion problem in personality research. Chicago: University of ChicagoPress; 1952.

Correspondence Addressed to: Luiz Pasquali
Campus Darci Ribeiro, ICC Sul - LabPAM, sala AI-096
Plano Piloto - Asa Norte
CEP 70910-900- Brasília, DF, Brazil