RELIABILITY AND VALIDITY

Gerardo Prieto and Ana R. Delgado
University of Salamanca

This article sets out to describe conceptually the psychometric properties of reliability and validity and the procedures used for assessing them. The part devoted to reliability, or test score accuracy, focuses on the models, procedures and statistical indicators most widely employed. As regards validity, the most important psychometric property, and that whose conception has changed the most, we summarize its history in testing contexts. The reader is warned that reliability and validity are not, as often thought, properties of the testing instruments, but rather of the particular inferences made from the scores. Another common error is to consider reliability and validity, not as questions of degree, but as absolute properties.

***Key words**: Reliability, Psychometrics, Testing, Validity.*

En este capítulo se describen conceptualmente las propiedades psicométricas de fiabilidad y validez y los procedimientos para evaluarlas. El apartado dedicado a la fiabilidad o precisión de las puntuaciones de las pruebas describe los distintos modelos, procedimientos empíricos e índices estadísticos para cuantificarla. En cuanto a la validez, la propiedad psicométrica más importante y la que ha experimentado mayores transformaciones a lo largo de la historia de la Psicometría, se resumen las principales concepciones y los debates en torno a la misma. Se previene al lector de dos frecuentes malentendidos: en primer lugar, considerar que la fiabilidad y la validez son características de los tests cuando corresponden a propiedades de las interpretaciones, inferencias o usos específicos de las medidas que esos tests proporcionan; en segundo lugar, tratar la fiabilidad y la validez como propiedades que se poseen o no en lugar de entenderlas como una cuestión de grado.

***Palabras clave**: Fiabilidad, Psicometría, Tests, Validez.*

Psychologists use diverse standardized procedures to obtain samples of people's behaviour. These resources, generically referred to as tests, include a scoring procedure providing measures that can be used for different purposes: to estimate a person's level in a construct (anxiety, quality of life, spatial visualization, etc.), to assess competence after a period of learning, to classify patients in diagnostic categories, or to select the most suitable candidates for a job. The legitimacy and efficiency of these practices depends on their reliability and validity.

Here we describe conceptually these two psychometric characteristics and the procedures most widely used for assessing them. Before going on, we should warn the reader about two common misunderstandings. The first concerns the notion that reliability and validity are characteristics of tests, when in fact they are properties of the interpretations, inferences or specific uses of the measures provided by tests. The second refers to the idea that reliability and validity are "all or nothing" qualities, that they are possessed or not, rather than being understood in terms of degree (AERA, APA & NCME, 1999).

RELIABILITY

Reliability is conceived as the consistency or stability of the measures when the measurement process is repeated. For example, if the weight readings of a basket of apples vary a great deal in successive measurements carried out in the same conditions, the measures will be considered unstable, inconsistent and unreliable. The lack of accuracy may have undesirable consequences for the cost of this product on a particular occasion. From this conception it follows that from the variability of the scores obtained in the repetitions of the measurement we can obtain an indicator of the reliability, consistency or accuracy of the measures. If the variability of the measures of the object is large, it will be considered that the values are inaccurate, and consequently, unreliable. Similarly, if a person were to do a test repeatedly *in the same conditions*, from the variability of the scores we could obtain an indicator of its degree of reliability. The impossibility of ensuring that the measurements are carried out in *exactly* the same conditions is one of the problems of psychological and educational measurement. A person's level of attention and motivation may vary as they repeat the same test over and over again, the difficulty of two supposedly identical tests designed for measuring the same construct may be unequal, the samples of examiners who mark a university admissions exam may differ in

Correspondence: Gerardo Prieto. *Facultad de Psicología. Universidad de Salamanca. Avda. De la Merced 109-131, 37005 Salamanca. España. Email: gprieto@usal.es*

their degree of strictness, and so on. Therefore, the efforts of assessors should focus on standardizing the measurement procedure to minimize the influence of those extraneous variables that can produce unwanted inconsistencies. Standardization of the procedure involves obtaining the measures on all occasions in very similar conditions: with the same test duration, the same instructions, the same practice examples, tasks with equivalent content and difficulty, similar qualification criteria for examiners, and so on.

The study of reliability starts out from the idea that the observed score in a test is a specific value of a random variable consisting of all the scores that could possibly have been obtained by a person in repetitions of the measurement process in similar conditions (Haertel, 2006). Obviously, it is not possible to repeat the measurement a very large number of times for the same participants. Therefore, the distribution of the scores is hypothetical, and its properties must be estimated indirectly. The mean of this distribution, which would reflect a person's level in the attribute of interest, is called *true score* in Classical Test Theory (CTT). CTT is an articulated set of psychometric procedures basically developed in the first half of the twentieth century, which has been used extensively for the construction, analysis and application of psychological and educational tests. Although CTT emerged in the context of the measurement of human aptitudes, its proposals extend to other areas. It is assumed that a person's true score does not change from occasion to occasion, so that the variability of the observed scores is due to the influence of a random, non-systematic *measurement error* (produced by factors which are unknown and uncontrollable in that situation). The amount of error in each case would be the difference between an observed score and the true score. The standard deviation of the errors, called *standard error of measurement* (SEM), indicates the accuracy of a person's scores, that is, their variability around the true score. The SEM reflects the error we can expect in an observed score. For example, if the standard error of measurement of an object's weight were 2 grams, it could be ventured that the observed weight will differ from the true weight by more than 2 grams only one third of the time. Although CTT permits us to estimate the SEM for people situated in different ranges of the variable (called *conditional* standard errors of measurement), it is customary to employ a single value applicable in a general way to all the scores of the people in a population. Obviously, appraisal of the SEM depends on the nature of the objects being measured: two grams is a negligible error if we are weighing very heavy objects such as sacks of cereals, but it is a crucial one in the case of lighter objects such as diamonds. That is, the value of the SEM is in the same units as the objects measured, and has no standardized upper limit that facilitates its appraisal. Hence the proposal of a standardized index of consistency or accuracy called *reliability coefficient*, which ranges between 0 and 1. From CTT it is derived that this coefficient is the ratio between the variance of the true scores and the variance of the observed scores in a population. Consequently, it indicates the proportion of the variability of the observed scores that cannot be attributed to measurement error; for example, if the reliability coefficient is 0.80, it is considered that 20% of the observed variability is spurious.

To estimate reliability statistics (SEM and reliability coefficient) empirically, various data-gathering designs are used that reflect different repetitions of the measurement process. The most well-known are called *test-retest* (application of a test to a sample of persons on two occasions between which the attribute remains stable), *parallel forms* (application to a sample of persons on the same occasion or on different occasions of two versions of the test equivalent in content, difficulty, etc.), *consistency between the parts of a test* (division of the test in two equivalent subsets of items or estimation based on the covariances between test items) and *consistency of the scores of different raters* (evaluation of a sample of behaviour by independent raters). Estimation of the reliability coefficient from these designs is usually based on the correlation between the observed scores obtained in the different forms of replication. An extensive literature provides detailed information on these procedures and on the concepts and developments of CTT. Excellent accounts can be found in the present issue (Muñiz, 2010) and in Gulliksen (1950), Martínez-Arias, Hernández-Lloreda and Hernández-Lloreda (2006) and Muñiz (1998).

In addition to CTT, other approaches are employed to quantify the reliability of test scores: Generalizability Theory (GT) and Item Response Theory (IRT).

CTT permits the quantification of just two components of the variance of observed scores: true variance and error variance. GT, conceived as an extension of CTT, attempts to specify the contribution to observed variance of a greater number of facets: variability between people, the number of times measured, different forms of the instrument, different raters, and interactions between components. The estimation of these influences takes place by means of variance analysis. The components distinct from the differences between persons (forms of the test, raters, measurement occasions, etc.) are interpreted as sources of error of the measurements, serving as evidence of possible causes of the error and permitting improvement of the measurement procedures. This model is especially useful for evaluating the reliability of the ratings given by raters to the products obtained in *open* texts or exams (examinees are not constrained by a closed format, as they are in multiple-choice tests). These matters are dealt with more fully in Brennan (2001), and in the article by Martínez-Arias (2010) in this issue.

IRT is a set of measurement models for estimating statistically the parameters of persons and items on a latent continuum on the basis of observable responses. In all statistical estimation procedures, the amount of estimation error is quantified on the basis of standard error (an index of the variability of the estimators of the parameter). The greater the standard error, the less accurate the estimation will be, and the greater the uncertainty about the value of the parameter. Likewise, in IRT models uncertainty about the location of a person or item in the latent variable is quantified on the basis of the *standard error of the estimate* (SEE) of the person or item. This statistic is distinct from the standard error of measurement of persons in CTT. As already described, SEM is a *global* measure of error, a single value applicable in a general way to all the scores of those in a population, which tends to underestimate or overestimate the degree of error affecting the scores situated at different levels of the variable. In contrast, SEE varies throughout the variable. Therefore, it can be considered an *individual* measure of accuracy, since it indicates the magnitude of the error used for estimating the parameters of the persons or items situated in different positions of the latent continuum. The function that describes how the SEE values of persons at different levels of the variable change is especially useful for determining the ranges in which a test is more reliable and for determining the reliability of the cut-off points employed in the classification of persons in diagnostic or performance-related categories.

Given that the SEE permits us to quantify an interval to estimate a person's parameter, the greater the interval, the greater will be the uncertainty about their location. If we take the opposite perspective, i.e., looking at how much *certainty* there is about a person's location, then we quantify the so-called information function, which is analogous to the reciprocal of the conditional error variance of CTT. The information function of a test indicates the extent to which the test permits us to distinguish between persons at different levels of the attribute. For a more detailed account, see Ayala (2009).

Before concluding this section we should refer to some practical considerations about the interpretation and use of reliability statistics, beginning by responding to one of the questions most frequently asked by test users: What degree of reliability should scores have for their use to be acceptable? Undoubtedly, the required magnitude depends on the consequences of the use of the scores. When the scores are to be used for making decisions with relevant consequences for the persons in question (e.g., acceptance or rejection in personnel selection), the reliability coefficient should be very high (at least 0.90). However, if it is a question of describing individual differences at the group level, it will be sufficient to attain more modest values (at least 0.70). Nevertheless, these conventions should be followed with caution: if the evaluation of the reliability has been carried out using procedures derived from CTT, the results will not necessarily be interchangeable, since the different data-collection designs mentioned previously (test-retest, parallel forms, internal consistency, etc.) detect different sources of error: instability of the measures, lack of equivalence of the tests, heterogeneity of the items, scarce agreement between raters, etc. Therefore, it is advisable to have reliability estimations based on different designs to achieve a better understanding of the error affecting the scores (Prieto & Muñiz, 2000). Moreover, the reliability statistics vary between populations and are affected by other conditions, such as length of the text and variability of the samples of persons. Consequently, what must be avoided is the error of considering that the estimation of the reliability based on a single study reflects the true and only reliability of the test. Test developers and users should provide detailed information on the quantification methods, the sample characteristics and the conditions in which the data were obtained (AERA, APA & NCME, 1999). As pointed out previously, the standard error of measurement is expressed in the same units as the test scores. Therefore, it is difficult to make comparisons between the reliability of the scores of different tests based on this statistic. In contrast, the magnitude of the reliability coefficient always ranges between standardized limits

(0 and 1), so that it is very useful for choosing the most reliable test among those potentially usable for a specific application. However, the standard error of measurement contributes more information for describing the accuracy of scores.

Sometimes, test scores are used not simply to estimate a person's position in the population of interest (referred to as *relative* interpretation), but to assign them to a diagnostic or performance-related category (pathological/normal, suitable/not suitable, accepted/excluded, etc.). To make this *absolute* type of interpretation it is customary to use cut-off points that guide the classification. Given that the reliability of the scores is not usually the same at all levels of the variable, it is important to know the degree of error around the cut-off point, since if it is high there will be a large number of false positives and negatives in the classification. In this case it is advisable to use the estimation error function or information function derived from IRT models.

We conclude this section by analyzing the relationship between score reliability and score validity, the property described in the next section. It is currently considered that validity, defined as the degree to which the interpretations and uses made of the scores are scientifically justified, is the most important psychometric property. Obviously, the utility of scores with scarce reliability for such purposes would be seriously compromised. Hence the consideration of reliability as a necessary condition of validity. However, it will not be a sufficient condition if the true scores, even though estimated in highly precise fashion, are not appropriate for achieving the objective of the measurement (representing a construct, predicting a criterion of interest, etc.). It is useful to bear in mind that reliability is a matter related to the quality of the data, whilst validity refers to the quality of the inference (Zumbo, 2007).

VALIDITY

The concept of validity has undergone substantial transformations over the past century, brought about by the great diversity of purposes for which tests have been used. According to Kane (2006), between 1920 and 1950 the principal use of tests consisted in predicting some variable of interest called *criterion* (for example, job or academic performance). Today, this approach continues to be of great importance when tests are used for selecting the most suitable candidates for a job, in admissions programmes, in the assignment of patients to treatments, etc. In such cases, assessment of a test's utility is usually quantified by means of the correlation between its scores and those of some criterion measure (*validity coefficient*). However, the success of this type of justification depends on the quality of the criterion measure, and especially of its representativeness (for example, are the indicators for measuring the criterion sufficient and representative of the job in question?). Hence the change of emphasis to a justification of the score in the criterion as coming from a sample of indicators that adequately represents the domain or *content* to be measured (the sum total of all possible indicators). This initial phase of development of the concept concluded, then, with the proposal of two excellent ways of establishing the validity of tests: criterion validation (correlation between the test scores and the scores in the criterion) and content validation (justification of the items for measuring the criterion as a representative sample of the content to be assessed).

Content validation extended from analysis of the criterion to that of the validity of the predictive tests: a test cannot be considered valid if the items making it up do not adequately sample the content to be assessed. Content validation is an especially fertile approach when the facets of the domain to be measured can be clearly identified and defined. This is the case of tests aimed at evaluating academic performance that can be specified according to the objectives of the instruction (concepts to be mastered and abilities to be possessed by a student). The methodology of validation rests fundamentally on experts' assessment of the pertinence and sufficiency of the items, as well as of the adequacy of other characteristics of the test, such as the instructions or the time it takes to complete. However, the precise specification of the content of the manifestations of constructs such as extraversion, working memory or achievement motivation is a more difficult task. Therefore, both content validation and criterion validation are considered insufficient for justifying the use of tests for evaluating cognitive aptitudes or personality attributes. Such dissatisfaction was behind an influential article by Cronbach and Meehl (1955), in which they proposed validation of the *construct* as the principal form of validation. As Cronbach (1971) pointed out, in a test for measuring a personality trait there is not simply one relevant criterion to predict, nor one aspect of content to sample. What there is, actually, is a theory about the trait and its relationships with other constructs and variables. If it is hypothesized that the test score is a valid reflection of the attribute, the assumption can be tested by analyzing its relationships with other variables. Consequently, construct validation can be conceived as a particular case of the testing of scientific theories by means of the hypothetico-deductive method. Although the user, in general, is not aware of it, measurement techniques imply theories (which are assumed to be sufficiently corroborated at the time of using them to test scientific or practical hypotheses), so that they themselves should be backed up by theories whose degree of sophistication will depend on the situation at that time of the research programme from which they emerged (Delgado & Prieto, 1997). Given that a theory postulates a network of relations between constructs and observable attributes, we cannot assume that they are valid if the theory is formally incorrect, the predictions derived from the theory are not met in the empirical data, or other, auxiliary assumptions have been violated. Thus, since the end of the 1990s a conception of construct validity has been imposed whereby it constitutes an integral framework for obtaining proof of validity, including that proceeding from criterion and content validation (Messick, 1989). The validation framework is defined on the basis of theories which specify the meaning of the construct to be evaluated, its relations with other constructs, its manifestations and its potential applications and interpretations. In addition to the tests necessary for ensuring an adequate representation of the construct, Messick included in the validation framework the justification of the consequences of the use of tests (individual and social implications). As we shall discuss later, the inclusion of so-called validation of the consequences is still the object of debate. This brief summary of the history of the validity concept, in which we have mentioned some important milestones, permits the understanding of the current concepts of validity and validation, whose principal characteristics we shall now describe.

Currently, *validity* is considered to refer to the extent to which empirical evidence and theory support the interpretation of test scores related to a specific use (AERA, APA & NCME, 1999). *Validation* is a process of accumulation of proof to support the interpretation and use of the scores. Hence, the object of validation is not the test, but rather the interpretation of its scores in relation to a specific objective or use. The validation process is conceived as an *argument* that starts out from an explicit definition of the interpretations proposed, from their theoretical foundations, from the predictions derived and from the data that scientifically support their pertinence. Since the predictions are usually multiple, a single proof cannot support a favourable judgement about the validity of the interpretations proposed. Multiple and convergent proofs are necessary, obtained in different studies. Therefore, validation is considered to be a dynamic and open process. Obviously, the related uses and interpretations can be quite varied. Thus, the sources of validation are multiple and their importance varies according to the objectives. The *Standards for educational and psychological testing* (AERA, APA & NCME, 1999) refer to the most important: test content, response processes, internal structure of the test, relations with other variables, and consequences of the use for which they are proposed. Before summarizing these methodological approaches, we should point out that they reflect different facets of validity, which brings them together as a single integrating concept. Therefore, it is not rigorous to use terms – *predictive validity*, *content validity*, *factorial validity*, and so on – that would lead to the notion of different types of validity.

Validation of test content

Tests are made up of a set of items designed to obtain a score representing a person's level in a construct (extraversion, maths ability, etc.). It would be difficult to guarantee the quality of the measures if the items did not sufficiently represent the different facets of the manifestations of the construct. In such a case the construct would be *under-represented*, so that the scores did not attain the required degree of validity. Likewise, there is evidence that responses to the items are influenced by variables extraneous to the construct of interest, and this constitutes one of the principal threats to validity, producing so-called *variance irrelevant to the construct*. Also among the objects of content validity are instructions, practice examples, test material, application time, and so on. The consultation of experts is the most usual way of appreciating the quality of the content, especially in educational contexts, though there is ever-increasing use of qualitative methods based on direct observation, interviews or file analysis. Standardized consultation procedures make it easier to obtain quantitative data indicating the percentage of quality items, the percentage of facets of the domain sufficiently evaluated, the percentage of judges rating the quality of the materials positively, inter-expert agreement, and so on. An exhaustive account of content validation can be found in Sireci (1998).

Analysis of response processes

Due to the influence of cognitive science, the validation of tests of intelligence, aptitude and performance must include analysis of the processes, the problem-solving strategies and the mental representations participants use for resolving the items. Evidence of validity will be obtained when the processes used fit with those postulated in the theories related to the construct measured. Study methodology is highly diverse: interviews with examinees in which they describe how they resolve the tasks, analysis of eye movements or response times, etc. When the theories about the construct have surpassed the merely exploratory stages, tests can be constructed on the basis of a *cognitive design* which specifies certain subsets of items to elicit certain latent processes. Responses to the items permit the estimation, through complex IRT models, of the person's parameters in the different cognitive components of the task and the identification of *types* of people who use different processing strategies. It is on this approach that the most advanced trends in cognitive diagnosis are based (Yang & Embretson, 2007).

Analysis of the internal structure of the test

Some tests provide a measure of a single construct, while others assess several constructs, including a subscale for each of them. Analysis of the internal structure is aimed at verifying empirically whether the items fit the dimensionality envisaged by the test developer. When attempting to adapt a test constructed initially for assessing persons in a specific population to a different population (from another culture, for example), it is obligatory to analyze whether the internal structure of the test remains invariant. If this is not the case, the meaning of the scores will differ between the two populations. The analysis of the internal structure of the test is generally carried out with the help of factor analysis models, which are described in detail in the article by Ferrando and Anguiano (2010) in this special issue.

Among the methods for evaluating the unidimensionality of the test, one of the most important is the analysis of *differential item functioning* (DIF). It can be stated that a test has similar validity in groups of different sex, culture, native language, etc. if its items do not present DIF, as dealt with in the article by Gómez-Benito, Hidalgo and Guilera (2010).

Association of the scores with other variables

The relationships between the test scores and other variables external to the test constitute an important source of validation. When scores are employed for selecting the most suitable candidates for a job, in admissions programmes, in the assignment of patients to treatments, etc., the justification is based on their utility for predicting an external criterion. The criterion is a measure of the variable of interest: job performance, presence or absence of a neuropsychological disorder, academic qualifications, etc. The utility of the test tends to be quantified by means of the correlation between its scores and those of some measure of the criterion (*validity coefficient*), or through other procedures: difference in scores between groups with different levels in the criterion, degree of agreement in the classifications into diagnostic categories made by means of the test and by experts, etc. The selection of a reliable and valid criterion (sufficient, objective and representative of the behaviour of interest) is the critical point that determines the goodness of the validation process. Depending on the point in time at which the criterion is evaluated, we can distinguish different types of data collection: *retrospective* (the criterion has been obtained before application of the test, e.g., based on a previous clinical diagnosis), *concurrent* (test and criterion scores are obtained in the same session) and *predictive* (the criterion is measured at a later point). The results may differ among these procedures: the preference will be for that which is most suited to the intended use (e.g., the predictive approach is the most appropriate for prognosis about future job performance). It is crucially important to analyze whether the predictive or diagnostic utility remains invariant across different groups of people. The question of the variability of the results for different groups, different studies, different criterion measures, etc. affects the generalization of the test's validity. Meta-analysis (see the article by Sánchez-Meca & Botella, 2010) allows the investigation of how correlations between the test and the criterion vary as a function of different facets of the studies.

When test scores are used for estimating people's levels in a construct, their correlations with those of other tests measuring the same or other constructs are of special relevance. The association between tests measuring the same construct is expected to be greater (*convergent validation*) than that between tests which measure different constructs (*discriminant validation*). To obtain empirical evidence, researchers make use of techniques such as factor analysis or the multitrait-multimethod matrix (Campbell & Fiske, 1959), which summarizes the correlations of a test with *markers* (tests of confirmed validity) that measure several constructs by means of different methods.

Validation of the consequences of test use

The latest version of *Standards for educational and psychological testing* (AERA, APA & NCME, 1999) considers the forecast of the possible consequences of test use as part of the validation process. From this perspective, the analysis and justification of the consequences are crucial elements when tests are to be employed for making critical decisions for people and institutions: selection, hiring, graduation, professional promotion, programme evaluation, and so on. The psychometric literature refers to such uses as high-risk. These practices are familiar ones within the Spanish context: selection of candidates for the job of pilot, army recruitment, security guard selection, public exams for entry to various institutions and companies, university exams, university entrance exams, evaluation of university staff, assessment of degree of dependence, authorization of arms licences and driving licences, and so on. In such cases, consideration of the pertinence of test use is not confined to whether or not the scores adequately represent the constructs or to the theoretical justification of the nomological network linking the constructs with the criteria of interest. High-risk applications have collateral effects of a personal and social nature. We can cite as an example of the former the effect on score validity of the training and learning of tests in which many of those entering selection programmes become involved. How sensitive are tests to this type of manipulation? There are other effects of an institutional nature, such as the peculiarity of the use of tests in a social context. Just consider the social fraud related to the use of the psychotechnical tests employed in our country for authorization of arms or driving licences. If we think of the consequences, could we say that they fulfil their function? Clearly, if validity refers to the extent to which the theory and the empirical evidence support the interpretation of test scores in relation to a specific use, the consequences can never be irrelevant to the validation process.

Although there appears to be some consensus on this matter, there are also discordant voices. For example, Borsboom and Mellenberg (2007) consider that the concept of validity should be more restricted in scope than it would be in the broad definitions proposed by Messick (1989) and the current version of *Standards*. In their view, validation should be confined to confirming whether there is a causal relationship between the construct and the test scores; interpretations of scores in applied contexts (personnel selection, accreditation, etc.) and the social impact of test use would strictly speaking lie outside the ambit of validity. While this simplified position may appear problem-free, defining construct validity as the validity of the causal inference implies identifying it with the internal validity of the evidence in favour of the construct (for an updated version of the different types of validity in experimental designs, see Shadish, Cook, & Campbell, 2002). Such identification could possibly be justified in well-advanced basic research programmes, but in practice it would render impossible the majority of psychological applications, not to mention the well-known problems of the concept of causation. Hence, pragmatism leads us to prefer a more flexible position, one which considers that validation procedures should serve to support the inference of the best possible explanation, including the evidence provided by the various qualitative and quantitative methods available to psychometricians at a given moment (Zumbo, 2007). If validation is considered to be a process open in time, validity is necessarily a question of degree, as suggested in *Standards*, and indeed, this conception of it is common to the different concepts of validity used by epistemologists.

The debate on the inclusion of the consequences in the concept of validity is not a technical question with which only the high-flying theorists of psychometrics are concerned. Advocating their inclusion brings with it responsibilities: can and should the developers of tests speculate on the desirable and undesirable consequences of their use? Which methodological repertoire should be used for it? Which authority should be responsible for the analysis and justification of the consequences? These and related questions will continue to feed the debate and generate proposals. An excellent review on the validation of consequences can be found in Padilla, Gómez, Hidalgo and Muñiz (2007).

Before concluding, we should refer to a terminological issue. In the tradition of test use in English-speaking contexts, *validation* has a legal meaning: "to declare legally valid". In contrast, in the Spanish language the term *validación* has two meanings: "the action and effect of validating", which it shares with the English word, and "the firmness, strength, safety or subsistence of an act". Although we tend to refer to the first meaning, the more aseptic one, it is in fact the second which comes closer to the objective of psychological research in its psychometric version.

REFERENCES

American Educational Research Association, American Psychological Association and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Borsboom, D., & Mellenberg, G.J. (2007). Test Validity in Cognitive Assessment. In J.P. Leighton and M.J. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education* (pp. 85-115). Cambridge: Cambridge University Press.

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer-Verlag.

de Ayala, R.J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 56, 81-105.

Cronbach, L. J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin,* 52, 281-302.

Delgado, A.R., & Prieto, G. (1997). *Introducción a los métodos de investigación de la psicología* [*Introduction to research methods in psychology*]. Madrid: Pirámide.

Gómez-Benito, J., Hidalgo, M.D., & Guilera, G. (2010). El sesgo de los instrumentos de medición. Tests justos [The bias of measurement instruments. Fair tests]. *Papeles del Psicólogo,* 31(1), 75-84.

Gulliksen, H. (1950). Theory of mental tests. New York, Wiley. Haertel, E. H. (2006). Reliability. In R.L. Brennan (Ed.), *Educational Measurement* (pp. 65-110). Westport, CT: American Council on Education and Praeger Publishers.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Westport, CT: American Council on Education and Praeger Publishers.

Martínez-Arias, M.R. (2010), Evaluación del desempeño [Performance assessment]. *Papeles del Psicólogo,* 31(1), 85-96.

Martinez-Arias, M.R., Hernández-Lloreda, M.J., & Hernández-Lloreda, M.V. (2006). *Psicometría* [*Psychometrics*]. Madrid: Alianza Editorial.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp- 13-103). New York: American Council on Education.

Muñiz, J. (1998). *Teoría Clásica de los Tests* [*Classical Test Theory*]. Madrid: Pirámide.

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems [Test theory: Classical Theory and Item Response Theory]. *Papeles del Psicólogo,* 31(1), 57-66.

Padilla, J.L., Gómez, J., Hidalgo, M.D., & Muñiz, J. (2007). Esquema conceptual y procedimientos para analizar la validez de las consecuencias del uso de los tests [Conceptual scheme and procedures for analying the validity of the consequences of test use]. *Psicothema,* 19, 173-178.

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.

Sánchez-Meca, J., & Botella, J. (2010). Revisiones sistemáticas y meta-análisis: herramientas para la práctica profesional [Systematic reviews and meta-analyses: tools for professional practice]. *Papeles del Psicólogo,* 31(1), 7-17.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Sireci, S.G. (1998). The construct of content validity. In Zumbo, B.D. (Ed.), *Validity Theory and the Methods Used in Validation: Perspectives From the Social and Behavioral Sciences* (pp. 83-117). Kluwer Academic Press, The Netherlands.

Yang, X., & Embretson, S.E. (2007). Construct Validity and Cognitive Diagnostic Assessment. In J.P. Leighton and M.J. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education* (pp. 119-145). Cambridge: Cambridge University Press.

Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.), *Handbook of Statistics, Vol. 26: Psychometrics*, (pp. 45-79). Elsevier Science B.V.: The Netherlands.