Papeles del Psicólogo, 2012. Vol. 33(1), pp. 12-21
http://www.cop.es/papeles

TESTS PUBLISHED IN SPAIN: USES, CUSTOMS AND PENDING MATTERS

Paula Elosua
University of the Basque Country

The tasks of twenty-first-century psychometrics include the development of formal models and the study of the conditions that guarantee an appropriate use of tests. Progress made in both directions is linked to the guidelines edited by national and international organizations that try to provide the applied professional with the latest advances and reflections. Nevertheless, we can see the distance that separates both worlds. In this paper, we analyze professional practice as it is reflected in the manuals of the most widely used tests in Spain. We tackle the treatment of reliability, validity, score interpretation or adaptation, taking the comprehensive guidelines by the APA and the guidelines of the International Test Commission as the criteria to be followed. The results show the gap between uses and duties. Throughout the paper, we tried to study some of the reasons that could explain this gap in depth.
Key words: *Tests, Uses, Standards, Psychometrics.*

La psicometría del siglo XXI asume entre sus tareas el desarrollo de modelos formales y el estudio y salvaguarda de las condiciones que garantizan un uso adecuado de los tests. Los progresos en ambas direcciones se conjugan en las directrices redactadas por organizaciones nacionales e internacionales que intentan acercar los últimos avances y reflexiones al profesional. Sin embargo, constatamos la distancia que separa ambos mundos. En este trabajo analizamos la práctica profesional tal y como está reflejada en los manuales de los tests más utilizados en España. Abordamos el tratamiento de la fiabilidad, validez, interpretación de puntuaciones o adaptación tomando como criterio de fuerza las directrices conjuntas de la APA y las directrices de la Comisión Internacional de Tests. Los resultados muestran las lagunas entre los usos y los deberes. A lo largo del trabajo intentamos profundizar en algunas de las razones que puedan explicar esta brecha.
Palabras clave: *Tests, Usos, Directrices, Psicometría.*

USES AND CUSTOMS

Based on initiatives emerging from the Test Commissions of the European Federation of Psychologists' Associations (EFPA), the General Council of Spanish Psychology Associations (Consejo General de Colegios Oficiales de Psicólogos; CGCOP) and the International Test Commission (ITC), we witnessed a description and review process of the uses related to tests. Projects endorsed by these organizations include studies on the test-related attitudes of psychologists (Evers and col., 2011, Muñiz and Fernández-Hermida, 2010), the analysis of the conditions that favor correct assessments (Muñiz, 2010) or the updating of national versions of questionnaires for their review in manuals published in Europe (Bartram, 2011; Muñiz, and col., 2011). The conclusions drawn from this reflection about our praxis allow us to illustrate the current scenario regarding uses and customs, which is fundamental and the precursor of any proposal for improvement in relation to "duties".

Strategies aimed at the improvement in the use of tests defended in Spain (Muñiz, 2010) advocate training and restriction. The objectives fixed in the formative strategy include professional training, dissemination of information on the quality and characteristics of tests, and the development of guidelines. The restrictive strategy limits the use of tests to qualified personnel. Focusing on the first strategy, we find that: a) psychologists who participated in the study on attitudes toward tests (Muñiz and Fernández-Hermida, 2010) recognize that the training they received in the Psychology undergraduate degree may not be sufficient for the correct use of most tests, b) psychometric knowledge advances in such a way that the distance between theoretic and practical psychometrics is today greater than ever, and c) the guidelines that attempt to unite and expound on the methodological and social advances in test theory are adhesion documents that endorse these but that in many cases are not followed, despite their generative role in the improvement of tests.

Correspondence: Paula Elosua. *Universidad del País Vasco. Avda. Tolosa, 70. 20018. San Sebastián. España.*
E-mail:Paula.elosua@ehu.es

.....
This study was partially funded by the Spanish Ministry of Economy and Competitivity (PSI2011-30256), and by the University of the Basque Country (UPV/EHU- GIU 12/32

The development of psychometrics as an area of knowledge in charge of psychological measurement has been marked from its origins by the distinction between different psychometric levels of intervention, which are perfectly portrayed in the contents of journals such as Psychometrika, or the closest to the applied psychologist, Psicothema. The distance between the contents of both is undeniable; however, psychometric knowledge is alive and well in both. On many occasions, psychometric theory contributions do not reach the applied psychologist, who is naturally more interested in practical substantive questions far removed from formal problems. A review of the most widely used tests in Spain (Muñiz and col., 2011) makes it clear that applied psychometrics is constructed on well-established concepts and uses, although somewhat "archaic" from the perspective of psychometric-theory advances. The most common practices (the use of the Cronbach´s Alpha coefficient as a reliability estimator, the use of the exploratory factor analysis technique as internal structure evidence, the estimation of the relationships between the test and convergent measures by means of the Pearson correlation coefficient or the score transformation for scale construction) are based on techniques and procedures developed in the first decades of the 20th century (Spearman, 1904; Thurstone, 1932). Their generalized use by professionals has required the conjunction of two elements: training and having tools available that are simple to use. The first is covered by undergraduate studies in psychology, where basic psychometric knowledge is universally offered to psychology students. The second is linked to the development of friendly software that integrates the analysis modules necessary in the test construction process.

Nonetheless, in the last 100 years psychometrics has advanced, and it has done so in a double direction. The most powerful theoretical models have been developed with the Classic Test Theory (Item Response Theory; Hambleton and Swaninathan, 1985), and the importance of the use of tests in order to safeguard their properties has been studied in depth. Today, we are more conscious than ever of the consequences derived from inadequate test use (Messick, 1995).

International associations related to test use take on the task of designing training and intervention plans to bring both worlds together. This effort results in the organization of congresses, training courses, and above all, in the elaboration of professional and technical guidelines. Guidelines for the use of the AERA, APA y NCME (1999 sometimes known as the APA Standards) tests, guidelines for evaluation in work and organizational contexts (EFPA), or the International Test Commission guidelines related to test adaptation, or evaluation via internet (http://www.intestcom.org) are excellent examples of this endeavor. They define reference frameworks that should be followed without exception, in which the methodological advances of the greatest impact are included and ethical principles are recommended that guarantee the correct use of tests. However, their content is not always reflected in professional practice.

OBJECTIVES

In the context of the review of uses and customs, the objective of the present paper is to analyze practice as it is reflected in the documentation that is offered by the tests published in Spain, and contrast it to some of the developments that, although not of the latest generation, have had greater impact on psychometric theory/practice. To do this, and with a descriptive and formative aim, the documentation of the most widely used tests in Spain has been analyzed (Muñiz and Fernández-Hermida, 2010), and has been compared to a well-established and accepted model: the guidelines for the use of tests published by the AERA, APA and NCME (1999) whose latest review is currently under discussion. The reliability, validity, norms, administration and adaptation sections are approached from a conceptual perspective and a methodological approximation in which the most relevant guidelines are referred to and the way of making them operative is studied.

Reliability. The guidelines define reliability in terms of consistency and error; in fact, the generic title of the section that deals with this point is "Reliability and Measurement Errors" (guideline 2.1. *"For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors and standard errors of measurement or test information functions should be reported"*). This duality is not the norm in the analyzed manuals.

Since the appearance of Cronbach´s article (1951), the use of the alpha coefficient as an indicator of the internal consistency of scores is present in documentation that accompanies tests and in articles regarding test construction/adaptation. Its success can be justified in the following three points: it is easy to estimate, it is easy to interpret (Nunnally, 1978) and it is not complicated to

improve its value. Practically all the manuals analyzed provide information about this indicator, which has become the most well known statistic of the Classical Test Theory.

Measurement error. However, consistency is not the only concept related to measurement precision; the standard measurement error concept occupies a pre-eminent place in individual assessment. The standard error of measurement (SEM or SE) quantifies the random error surrounding a true score, and in assessment contexts where the final objective is the interpretation of a score, its relevance is greater than that of internal consistency; it offers a way of expressing uncertainty in relation to scores that is not offered by Cronbach´s alpha coefficient. Only from the SE value can we make statements such as "With a 95% probability, person X´s score lies between values 34 and 48".

The necessity of offering more information about the internal consistency and measurement error affects all and every one of the partial scales that measure well-differentiated behavioral areas or aspects within a test (Elosua, 2008).

The importance of informing about score (un)certainty increases in situations that demand the establishment of cut-off points in interpretation. Despite the fact that one of the basic principles of classic test theory assumes the consistency of the standard error throughout the score continuum, its compliance is normally violated (Hambleton and Swaminathan, 1985). If there are signs of said violation, and in addition the test offers score intervals with diagnostic or selection criteria, it is important to estimate the effect of the measurement error around the critical scores. Likewise, if different interpretation criteria are offered in the test as a function of gender, age or another relevant variable related to the test, a common situation in professional practice, it would be convenient to estimate the standard error of measurement in a differentiated manner for each one of the groups considered.

The estimation of the standard error of measurement from the classical theory is as easy as calculating Cronbach´s alpha coefficient; despite this, most available computer programs do not offer information on this index among its standard outputs.

Consistency from a model. Despite the extended use of the alpha coefficient as a reliability estimator, an increasing number of psychometrists advise against its use, and propose alternatives of estimation and definition of reliability constructed on alternative models of measurement to the classical test theory (McDonald, 1981,1999; Jöreskog, 1971; Raykov, 2001).

The new approximations to the study of score consistency are based on item response models or on factorial models that go deeper into the problem of homogeneity with the factor measured. The estimation of consistency from the factorial perspective would help to eliminate: a) the incorrect use of the alpha coefficient as an indicator of unidimensionality (Hattie, 1985), b) the problems derived from the use of statistics that do not meet the model assumptions in reference, in this case, to the continuous nature of variables or the tau-equivalence of measures (Zumbo, Gadermann and Zeisser, 2007), and c) they would allow us to study the homogeneity of measures from structural equation models in depth, for which estimation software such as Mplus (Muthén and Muthén, 2001), FACTOR (Lorenzo-Seva and Ferrando, 2007), LISREL (Jöreskog and Sörbom, 1996), EQS (Bentler, 1995), AMOS, or R (Ihaka and Gentleman, for codes see Elosua and Zumbo, 2007) may be used.

In the study of reliability, the models constructed from the Item Response Theory (IRT) deserve a special section given that they represent an important advancement with respect to classical test theory in the test construction and item analysis processes (Embretson and Reise, 2000; Lord, 1980; van der Linden and Hambleton, 1997). The item response models facilitate the study of: a) invariance in both individuals and items, b) group equivalence, and c) the conditional estimation of measurement errors (informative function). The advantages and formal properties of IRT make it an attractive and effective theoretical framework in the resolution of associated measurement problems in psychology, among these, score equivalence and comparison, differential item functioning analysis, construction of adaptive tests and the elaboration of evaluative reports.

Validity. If there is a point on which there is no discussion among theorists and professionals, it is that which makes reference to the importance of validity and the process of score validation in the construction and use of tests: *"Validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests"* (APA, AERA and NCME, 1999, page 9).

From a historical perspective, the definition and evolution of the validity concept has been reflected in successive publications of the joint guidelines of the APA

(1954, 1966, 1974, 1985, 1999; see Elosua, 2003). The differentiation between construct validity, predictive validity and content validity has impregnated the first four editions. For the first time, the 1985 edition defends a unitary conception of validity, although it distinguishes between three types of evidence. In the 1999 edition, validity is defined as a unitary concept, postulating five sources of evidence, and insists on the practical aspect of validity. The about-turn adopted implicates linking validity of scores to their use (a perspective that is maintained in the next edition). The guidelines suggest the use of five sources of evidence: content evidence, evidence based on internal structure, evidence based on the relationships with other variables, evidence regarding the response process and evidence based on the consequences of the test.

The new theory on validity (validation) places the focus on the validation of the proposed interpretation; we no longer talk of test validity and in that context it does not make sense to speak of content validity, criteria validity and construct validity.

The goal is to justify an interpretation of the scores based on reasons and arguments that are gathered during the validation process (Kane, 1992, 2006). This idea, which is included in the 1999 guidelines *("validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use"*. APA, AERA and NCME, 1999, page 9), seeks to justify the use of scores as a representation of the construct that is being measured, or whenever it is pertinent, justify their utility in the prediction of behaviors.

These definitions of validity offered in the 30s, which differentiated three types of validity, are the most commonly adopted in the analyzed manuals. They reflect the idea that a test is valid for that with which it is correlated (Kelley, 1927; Guilford, 1946), or they make test validity equal to the degree to which the test measures what it attempts to measure. The operationalization of the first meaning is extremely simple; it is enough to estimate the correlation between a test and a criterion; that is, it is enough to estimate the validity coefficient. The second meaning, which is more in accordance with or closer to the factorial models, is materialized normally by means of exploratory factorial analysis. Although both definitions have been accused of being incorrect for decades (Anastasi, 1954; Rulon, 1946), the procedures linked and derived from these are still valid in current practice. The factorial techniques and the correlation studies (regression) are present in all the analyzed manuals. They are techniques designed at the inception of psychometrics (Spearman, 1904, Thurstone, 1932) which have influenced the first definitions of validity, are integrated in all study plans of Spanish Psychology Faculties and are part of software modules for data analysis in social sciences.

However, just as the validity (validation) concept has evolved, the original factorial and regression models have led to more powerful and explicative methodologies designed for the study of the relationships between observed and latent variables: structural equation modeling (SEM). Since the decade of the 70s, a rapid development of SEM theoretical models and friendly software for its estimation has taken place (Bentler, 1980, 1986; Bollen, 1989; Jöreskog and Sörbom, 1996), which does not appear reflected in the manuals of the analyzed tests. Structural Equation Modeling represents a family of powerful and flexible multivariate statistical techniques, among which the factorial confirmatory models are included, which allow the modeling of the relationships between latent and indicative variables, assuming the presence of measurement errors in all cases. The regression and correlation models do not contemplate this fact. The applicability of the SEM techniques has been well documented both for correlation studies and for studies of an experimental nature. Their advantages in cross-sectional or longitudinal studies, among which the design of growth curves acquire special importance, is evident; they favor and impel the role of the theory in applied research and allow us to contrast and evaluate alternative explicative models (Kline, 2010; Millsap and Maydeu-Olivares, 2009). The advantages associated to the use of structural equation models in the validation process include reliability and measurement error estimation, the construction of explicative models and their simultaneous contrast for different groups.

The software for the application of the SEM models is varied; LISREL (Jöreskog and Sörbom, 1996), EQS (Bentler, 1995) AMOS, Mx (now integrated in R) or the *sem* and *lavaan* in R packages (Elosua, 2009; Ihaka and Gentleman, 1996).

Scales and score comparison. The transformation of raw scores into derived scores (standard scores, percentile ranks, graded equivalents), whose finality is to favor the interpretation or the definition of cut scores that discriminate between diagnostic categories or

performance levels and the establishment of selection criteria, occupy a specific section in the guidelines. In these, the importance of distinguishing between normative and criterial interpretations of the scores is highlighted. In the first case, which is the most prevalent in the analyzed manuals, the scores are read in reference to the statistical distribution of the scale sample; a score is interpreted in function of how the normative group has carried out the test and to do this, transformations in percentile ranks, z scores, T scores, grade equivalent scales or other types of derived scores are used. The criterial interpretation is different, and its use requires the definition of an external referent with respect to which the execution levels are compared. In this regard, it is possible to inform about the percentage of correct answers in a specific domain, the probability of answering an item correctly, or the probability of presenting a certain pathology or trait.

Theoretical distinction (norm-criterion) is not mutually exclusive, but its combined use has to be based on documentation that justifies one as well as the other. There are several manuals that create substantive categories in basis of normative distributions, transforming a percentile into a diagnostic category. Nevertheless, on this point the guidelines are clear (Guideline 4.9. *"When raw score or derived score scales are designed for criterion-referenced interpretation including the classification of examinees into separate categories, the rationale for recommended score interpretations should be clearly explained"*. Guideline 4.19. *"When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented"*.).

The adoption of a criterial interpretation of the scores requires that the method and procedures used for their determination be clearly specified (Cizek and Bunch, 2007; Hambleton and Pitoniak, 2006). This practice, which is common in educational assessment on a large scale or in international programs such as PISA or TIMMS, is not present in the test manuals.

The criterial reading of the results acquires a greater relevance when the objective of the evaluation is a differential diagnosis (Guideline 12.6 *"When differential diagnosis is needed the professional should choose, if possible, a test for which there is evidence of the test's ability to distinguished between the two or more diagnostic groups of concern rather than merely to distinguished abnormal cases from the general population"*). In the assessment contexts, the arguments offered in the validation process should gather information about the plausibility of the inferences regarding a diagnosis. The description of the arithmetic measures obtained in different groups is not a valid argument. The arguments should include confidence intervals, effect sizes, or tables showing the degree of overlap of the distributions among diagnostic samples, discriminant analyses or other techniques derived from the mining of data estimating classification/prediction functions (Bully and Elosua, 2011).

Administration, correction and reports. The administration, correction and elaboration of the resulting assessment report are unavoidable tasks in the establishment of standardized measurement processes. Only when the exam opportunities and conditions are equitable can we talk of standardized measures.

The study of optimal exam conditions does not exclude, but on the contrary, requires the consideration of accommodation measures for those evaluees who need them, whether for not possessing the necessary level of language dominance for a correct evaluation, or for the presence of motor or other types of disabilities. The accommodation methods that are common in educational assessment are not yet included in our praxis.

The importance of the adequate elaboration of assessment reports is being unanimously recognized by the psychometric community (Hattie, 2009; Hambleton and Zenisky, in press). The summative, diagnostic, and normative information must be drafted in an intelligible and clear manner for the final recipient. It must offer information about the quality of the measure and its interpretation according to the purpose of the test in an effective manner. In this regard, guideline 5.10 reminds us that *". …the interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used."*

Test Adaptation. The edition of the APA standards regarding test use does not have a section dedicated to test adaptation, a shortage that is covered by the guidelines elaborated by the International Test Commission. They are unsurpassable as a reference criterion. Their importance in the context we are dealing with is clear if we take into consideration that of the 10 most frequently used tests in Europe (Elosua and Iliescu, in press) 9 were constructed in the English language, and of the 25 most frequently used tests in Spain, 17 are adaptations of versions constructed in another language.

The guidelines for the adaptation of tests were published for the first time in the year 1994 (Muñiz and Hambleton, 1996), and in their second edition (Bartram, Gregoire, Hambleton, and van de Vijver, 2011; Elosua and Hambleton, 2011) they offered a model in which the points to consider in test adaptation are clearly described and operationalized. The 20 guidelines are structured in 6 categories with the aim of covering all the stages implicated in the adaptation process: previous guidelines, development guidelines, confirmation guidelines, administration guidelines, scoring guidelines and guidelines referring to documentation.

One of the most important guidelines and which can summarize the content of all these reminds us of the importance of offering empirical information about the construct equivalence, method equivalence and item equivalence in all populations to which the test is destined (second confirmation guideline). It synthesizes the relevance of the equivalence analysis between the object of measurement in different populations and the method used to measure it. The approximations of their study can be qualitative (procedures of judgment, content analysis, interviews) and quantitative. From the latter, classic exploratory approaches can be adopted, such as Tucker's congruence index (Tucker 1951); however, it is recommendable to use invariance models derived from item response models or structural equation models (Elosua and Muñiz, 2010).

PROCEDURE

The latest edition of a study about psychologists' opinions carried out in Spain (Muñiz and Fernández-Hermida, 2010) has revealed, along with other points, the tests most widely utilized in professional practice (Table 1). In the list of the 10 most widely used tests by specialty (Clinical, Educational and work) we find that of the 25 questionnaires, 24 are standardized measurement instruments. Of these, 15 were constructed in the United States, 1 in Great Britain, 1 in Switzerland, 1 in France, 1 in Italy, and 6 have a Spanish origin. These data reflect that 76% of the most widely used tests in Spain are adaptations, and that of these, 15 (79%) come from the United States.

The 10 most widely used tests in Spain, regardless of the professional specialty (see Table 1), are adaptations; of these, 9 were constructed originally in English and their first versions were published decades ago. Raven´s Progressive Matrices test is the oldest; its first edition was published in 1938. It is remarkable to confirm that the average year of publication of the top ten brings us to the year 1960, which is an indicator of the strength and current nature of these tests. Of these, some are constructed on solid formal models. Such is the case of the cognitive scales (WAIS, WISC) or Cattell´s personality measures that were constructed based on factorial models of intelligence or personality. Other scales have a more eclectic nature and emerge from practical or applied considerations, such as the MMPI, SCL or BDI.

The classification of the tests according to their psychological domain allows us to confirm that of the 25 tests, 12 are of a cognitive nature (WAIS, WISC, Raven, BADYG, TALE, MSCA, PROLEC, BENDER, ITPA, TAMAI, DAT, IGF), 6 are personality questionnaires (16PF, NEO PI-R, PAPI, TPT, IPV, BFQ) and 6 are of a clinical nature (MCMI, MMPI, SCL-90, BDI, STAI, MMSE).

Of the 25 tests that make up the universe of the study, 22 have been analyzed. The Rorschach test was excluded due to its projective nature; the BDI is a "phantom" questionnaire in the sense that although it appears on the list of the most widely used tests in professional practice, when this study was carried out there was no Spanish version of it; finally, the PAPI was not studied as we did not have access to it.

RESULTS

Reliability. The treatment granted to the reliability topic in the manuals mainly corresponds to a classical conception in which reliability is considered equal to internal consistency. In this regard, Cronbach´s alpha coefficient is the preferred indicator. A total of 15 manuals offer this information. The temporal stability of scores analyzed by means of a test-retest is approached by a total of 6 manuals. Unfortunately, information on standard errors of measurement is not present in all manuals; only 5 offer this information. If the convenience of offering information about standard errors conditional to the scores of scales is considered, this number decreases. One of the analyzed manuals provides information about the Sem conditional to the score in the framework of the Item Response Theory, and 4 tests estimate this statistic by sample.

Validity treatment. The perspective adopted by the majority of the analyzed manuals does not adapt to the theoretical framework defended in the AERA, APA and NCMEA (1999) standards. Even if we consider the threefold concept of validity (content, criteria, construct),

only 11 of the manuals make reference to the three forms of validity. Content evidence is included in 8 manuals; they inform about internal structure in 16 manuals and about the test relationships with other variables in 17. Only a brief reference to the response process or the consequences has been found in 1 manual.

Validation methodology. Most tests support their evidence with correlational studies, which are used in 18 manuals. The regression model is used in 2 manuals. Exploratory factorial analysis is estimated in 11 manuals, and information about the confirmatory models is offered in 6 manuals.

Score interpretation. In the score interpretation process most of the manuals analyzed propose normative interpretations based on the score distribution or grading norms. In addition, 7 of these offer criterial interpretations. Cut scores between categories are displayed in 14 manuals. However, justifications or evidence for the definition of these cut scores are only offered in 8 manuals. The manuals offer tables or scales for the interpretation of the scores. Non-lineal transformations are used in 20 manuals and lineal transformations in 19. The majority of manuals do not add any information based on the standard error in the reading of the results. The manuals offer tables or scales for the interpretation of scores.

Adaptation. Although most of the analyzed tests are adaptations, the adaptation process does not appear to be well documented. Only 8 tests discuss the problems related to language equivalence. Tucker´s congruence coefficient appears in 4 manuals. No evidence has been found of structural or metric equivalence studies based neither on SEM models nor on the study of differential item functioning.

DISCUSSION

Tests are the most well known aspect of psychometric research and have the greatest social impact. Since the beginning of the 20th century, they have served psychological research and have helped in the decision-taking processes in educational, social, legal or clinical fields. Throughout history, their utilization and consequent social consideration has gone through periods of apogee and crisis, in whose origin we find abusive and incorrect uses.

TABLE 1
MOST UTILIZED TESTS IN SPAIN

	Test Name	Field	1st edition	Country of origin	Adaptation
MCMI-III*	Milon Clinical Multiaxial Inventory	C-T	1997	USA	2007
16PF-5*	Sixteen Personality Factors Questionnaire-5	C-E-W	1949	USA	1995/2005
MMPI-2-RF*	Minnesota Multiphasic Personality Inventory-2-Restructured Form	C-W	1943	USA	2009
BDI-II*	Beck Depression Inventory-II	C	1961	USA	-
WISC-IV*	Wechsler Intelligence Test for Children-IV	C	1949	USA	2005
WAIS-III*	Wechsler Adult Intelligence Scale-III	C	1955	USA	1999
STAI*	State-Trait Anxiety Inventory	C	1970	USA	1994
RORSCHACH*	Rorschach test	C	1921	Switzerland	
SCL-90-R*	Symptom Checklist-90-Revised	C	1975	USA	2001
MMSE	Mini-mental State Examination	C	1975	USA	2002
BADYG	Bateria de Aptitudes Diferenciales y Generales	E	1989	Spain	-
TALE	Test de Análisis de Lecto-Escritura	E	1980	Spain	
MSCA	McCarthy Scales of Children´s Abilities	E	1972	USA	1977/2006
RAVEN*	Raven´s Progressive Matrices	E	1938	Great Britain	2001
PROLEC-R	Batería de Evaluación de los Procesos Lecotres Revisada	E	1996	Spain	-
BENDER	Bender´s Visual-Motor Gestalt Test	E	1938	USA	1993
ITPA	Illinois Test of Psycholinguistic Abilities	E	1968	USA	1984/2004
TAMAI	Test Autoevaluativo Multifactorial de Adaptación Infantil	E	1983	Spain	No
PAPI	Personality and Preference Inventory	W	1960	USA	
DAT-5	Differential Aptitude Tests	W	1947	USA	1960/2002
TPT	Test de Personalidad de TEA	W	2002	Spain	-
IPV	Inventaire de Personnalité des Vendeurs	W	1977	France	2005
IGF	Inteligencia General y Factorial Renovado	W	1991	Spain	-
BFQ	Big Five Questionanire	W	1993	Italy	1995/2007
NEO-PI-R	NEO Personality Inventory Revised	W	1978	USA	1999/2008

(C-Clinical, E-Educational, W-Work: * 10 most used regardless of specialty)

In professional practice, psychological tests are support tools in diagnosis, the design of intervention and assessment plans, and in professional selection. Regardless of the test modality (cognitive, neuropsychological, adaptive, social, behavioral, personality or vocational), and of its purpose, it is an instrument that must be constructed following principles that guarantee its technical quality (Wilson, 2005), and must be used according to criteria that allow these to be safeguarded. Only an appropriate use conforming to the purposes for which it was constructed will guarantee the validity of its interpretations.

National and international organizations related to test use try to improve professional practice using training as their main argument. Although the main training base of the professional is the undergraduate degree offered in our universities, this may not be sufficient (Muñiz and Fernández-Hermida, 2010). The content of the study plans are entrenched in the inertia marked by years of tradition and uses, and it feeds back with respect to established practices. However, psychometrics has evolved, and the distance between psychometric theory and professional practice is now greater than ever. The development of psychometrics is marked by the construction and study of new formal models, but also by awareness with respect to the social relevance of tests, which was previously non-existent.

The standards elaborated by professional organizations fulfill an important mission in the dissemination of formal and social advances. They are documents that combine rigor and simplicity in texts of easy reading and comprehension. They provide general norms, which are important in the process and evaluation of the result of the construction/adaptation and use of tests. In this article, which follows the educational and informative line started by the CGCOP, we attempted to show the reflection of two of the most important references related to test use, the comprehensive guidelines by the APA and the guidelines written by the International Test Commission. The result shows the uses and customs established in our practice, but also the matters pending.

The most important conclusions referring to each one of the points analyzed could be summarized as the need to provide information regarding standard error, the need to update the idea of validity toward a more dynamic and argumentative conception, the convenience of using confirmatory and explicative models in the study of the relationship between variables, the interest in justifying criterial interpretations and the use of adequate models for estimating the likeliness of a diagnosis, or the importance of guaranteeing equivalence in the adaptation of tests by means of invariance studies. None of the points is new in psychometric research and the work by Muñiz-Fernández-Hermida (2010) shows that professionals are aware of some of these deficiencies (e.g., not using standard error is recognized as a problem by professionals). However, after more than a decade of publishing guidelines, their content is not reflected in the documentation analyzed. The guidelines delineate a path to follow and the responsibility for the transition from customs to duties is shared by all; professionals, editors, professional colleges and professors.

FIGURE 1
APPROXIMATIONS USED IN THE ANALYZED MANUALS (N=22)

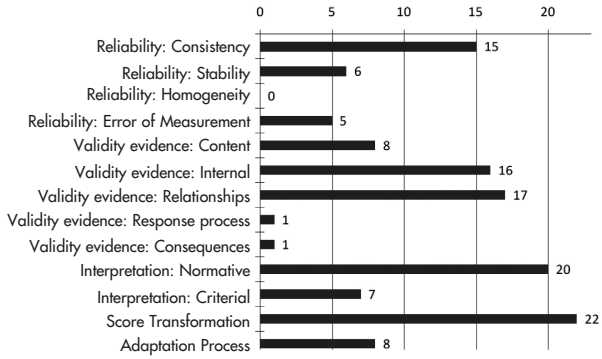
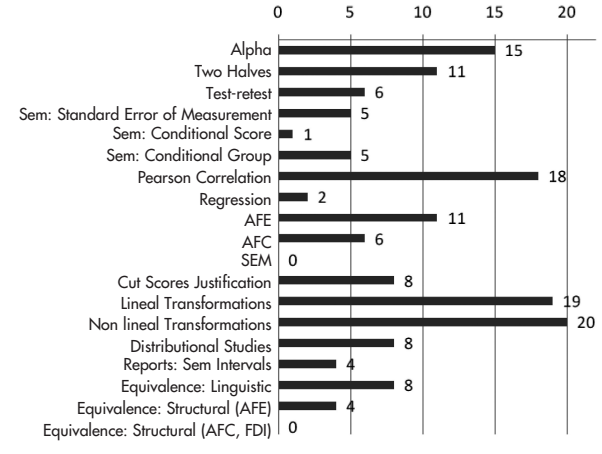


FIGURE 2
METHODOLOGY USED IN THE ANALYZED MANUALS (N=22)



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 Psychological Association.

Anastasi, A. (1954). *Psychological testing*. New York: Macmillan.

Bartram, D. (1998). The need for international guidelines on standards for test use: A review of European and international initiatives. *European Psychologist, 2*, 155-163.

Bartram, D. (2011, July). *The EFPA Test Review Model: Time for an Update?* Symposium organizado en el 12th European Congress of Psychology, Estambul, Turquía.

Bartram, D., Gregoire, J., Hambleton, R. K., and van de Vijver, F. (2011, Julio). *International Test Commission Guidelines for adapting educational and psychological tests (2nd edition)*. Sesión especial en el 7th Conference of the International Test Commission, Honk Kong, China.

Bentler, P. M. (1995). EQS structural equations program manual. Encino, CA: Multivariate Software Bentler, P.M. (1980). Multivariante analysis with latent variables: casual modeling. *Annual Review of Psycholy 31*, 419-456.

Bentler, P.M. (1986). Structural modeling and psychometrika: an historical perspective on growth and achievements, *Psychometrika, 51*, 31-51.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley

Bully, P. and Elosua, P. (2011, Julio). *Classification procedures and cut-score definition in psychological testing: A review*. Comunicación presentada en el 11th European Conference on Psychological Assessment. Ritga.

Cizek, G. J., and Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

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

Embretson, S. E. and Reise, S. P. (2000). Item Response Theory for psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J.R., Frans, O., Gintiliené, G., Hagemeister, C., Halama, P., Iliescu, D., Jaworowska, A., Jimenez, P., Manthouli, M., Matesic, K., Schittekatte, M., Sümer, C., and Urbánek, T. (in press). Testing practices in the 21th century. Developments and European Psychologist's opinions. *European Psychologist, 17*, 300-319.

Elosua, P. (2003). Sobre la validez de los tests. *Psicothema, 15(2)*, 315-321.

Elosua, P. (2008). Una aplicación de la estimación Bayes empírica para incrementar la fiabilidad de las puntuaciones parciales. *Psicothema, 20(3)*, 497-503.

Elosua, P. (2009). ¿Existe vida más allá del SPSS? Descubre R. *Psicothema, 21(4)*, 652-655.

Elosua, P. and Hambleton, R.K. (2011, Julio). *Nuevas directrices de la ITC para la adaptación de tests*. Trabajo presentado en el XII Congreso de Metodología de las Ciencias Sociales y de la Salud, San Sebastián.

Elosua, P. and Iliescu, D. (2011). *Psychological test validity. Where we are an where we should to go*. Comunicación presentada en el 12th European Congress of Psychology, Estambul, Turquía.

Elosua, P. and Iliescu, D. (2012). Test in Europe. Where we are and where we should to go. *International Journal of Testing, 12*, 157-175.

Elosua, P. and Muñiz, J. (2010). Exploring the factorial structure of the Self-Concept: A sequential approach using CFA, MIMIC and MACS models, across gender and two languages. *European Psychologist 15*, 58-67.

Elosua, P. and Zumbo, B. (2008). Reliability coefficients for ordinal response scales. *Psicothema, 20*, 896-901.

Guilford, J.P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-439.

Hambleton, R. K. and Swaminathan, H. (1985). *Item response theory. Principles and applications*. Boston, Kluwer: Nijhoff Publishing.

Hambleton, R. K. and Pitoniak, M. (2006). Setting performance standards. En R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.

Hambleton, R.K. and Zenisky, A. (en prensa). *Reporting Test Scores in More Meaningful Ways: A Research-Based Approach to Score Report Design*

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement 9*, 139–164.

Hattie, J. (2009, April). *Visibly learning from reports: The validity of score reports*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Ihaka, R., and Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5, 299–31

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109 133.

Jöreskog, K.G. and Sörbom, D. (1996). LISREL8. User's Reference Guide. Chicago:Sci Software Int. Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. T. (2006). Validation. En R. Brennan (Ed.), *Educational measurement, 4th ed* (pp. 17-64). Westport, CT: Praeger

Kelley T. L. (1927). *Interpretation of educational measurements*. Yonkers, NY, World Book Company.

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling (3rd Edition)*. New York: The Guilford Press.

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

Lorenzo-Seva, U. and Ferrando, P. J. (2007). *FACTOR: A computer program to fit the exploratory factor analysis model*. University Rovira y Virgilli

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology 34*,110–117.

McDonald, R. P. (1999). Test theory. A unified treatment. Mahwah, NJ, Lawrence Erlbaum Associates. Messick, S. (1995). Validity of psychological assessment. *American psychologist, 50,* 741–749.

Millsap, R. and Maydeu-Olivares, A. (Eds.) (2009). *Handbook of quantitative methods in Psychology*. London: Sage

Muñiz, J. (2010, Julio). *Estrategias para mejorar el uso de los tests*. Comunicación presentada en el Congreso Iberoamericano de Psicología, Oviedo

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

Muñiz, J. and Fernández-Hermida, J.R. (2010). La opinión de los psicólogos españoles sobre el uso de los tests. *Papeles del Psicólogo, 31*, 108-121.

Muñiz, J. and Fernández-Hermida, J.R., Fonseca-Pedrero, E., Campillo-Alvarez, A. y Peña-Suarez, E. (2011). Evaluación de tests editados en España. *Papeles del Psicólogo, 32*, 113-128.

Muñiz, J. and Hambleton, R.K. (1996). Directrices para la traducción y adaptación de tests. *Papeles del Psicólogo, 66*, 63-70.

Muthén, L. K. and Muthén, B. O. (2001). Mplus user's guide. Los Angeles: Muthén y Muthén. Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Raykov, T. (2001). Estimation of Congeneric Scale Reliability via Covariance Structure Analysis with Nonlinear, *British Journal of Mathematical and Statistical Psychology, 54,* 315-323.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16*, 290 296.

Spearman, C. (1904). "General intelligence," objectively determined and measured. American *Journal of Psychology 15*, 201-293.

Thurstone, L. L. (1924/1973). *The Nature of Intelligence*. London: Routledge.

Tucker, L.R. (1951). *A method for synthesis of factor analysis studies. Personnel Research Section Report, 984*. Washington, D. C.: Department of the Arm

Van der Linden, W. and Hambleton, R.K. (1997). *Handbook of modern ítem response theory*. New York: Springer Verlag.

Wilson, M. (2005). *Constructing Measures. An Item Response Modeling Approach*. Mahwah, NJ: Lawrence Erlbaum

Zumbo, B. D., Gadermann, A. M. and Zeisser, C. (2007). Ordinal Versions of Coefficients Alpha and Theta For Likert Rating Scales. *Journal of Modern Applied Statistical Methods, 6,* 21-29.