ISSN 1697-2600
2007, Vol. 7, Nº 3, pp. 863-882

International Journal of
IJCHP
Clinical and Health Psychology

Standards for the development and review of instrumental studies: Considerations about test selection in psychological research

Hugo Carretero-Dios[1] and Cristino Pérez *(University of Granada, Spain)*

(Recibido 7 de mayo 2007 / Received May 7, 2007)
(Aceptado 11 de junio 2007 / Accepted June 7, 2007)

ABSTRACT. This paper discusses the criteria that should be considered when selecting psychological assessment tests in a research context. Traditionally attention has focused – and still does – on the stages that must govern any process of test construction/adaptation. This work is guided by internationally accepted standards, whose scientific importance is agreed by the scientific community. However, beyond any construction/adaptation process, the use of tests is a constant feature of psychological research, so it is of vital importance to select the tests adequately. For this reason, in this theoretical study we provide a summary of the criteria that should guide test construction/adaptation as well as some general guidelines to consider when selecting tests for psychological research. The information presented is organized into six sections, each of which corresponds to a different stage in the process of test creation: a) conceptual definition of the construct to assess; b) information about item construction and qualitative assessment; c) results of the statistical analysis of the items; d) empirical evidence of the internal structure of the test; e) results of the reliability estimation; and f) external evidence of score validity. The study ends with a reflection on the scope of the proposed guidelines and the importance of using clear criteria to select the tests used in research.

KEY WORDS. Standards for the review of instrumental studies. Test construction. Test adaptation. Test selection. Theoretical study.

[1] Correspondence: Facultad de Psicología. Universidad de Granada. 18071 Granada (Spain). E-mail: hugocd@ugr.es

RESUMEN. En este trabajo se discuten los criterios a tener en cuenta a la hora de seleccionar tests de evaluación psicológica en un contexto de investigación. Tradicionalmente la atención se ha centrado y se centra sobre las fases que deben regir todo proceso de construcción/adaptación de tests. Estándares internacionalmente aceptados sirven para dirigir este trabajo, y la comunidad científica coincide en la importancia de éstos. No obstante, y más allá de cualquier proceso de construcción/adaptación, el hecho es que el uso de tests es una constante en la investigación psicológica, y una adecuada selección de las pruebas resulta un asunto de vital importancia. Por ello, y esquematizando en primer lugar los criterios que deben guiar la construcción/adaptación de tests, en este estudio teórico se desarrollan unas directrices generales a tener en cuenta a la hora de seleccionar tests para efectuar una investigación psicológica. La información va a presentarse organizada en un total de seis apartados, cada uno de los cuales corresponde a una fase distinta dentro del proceso de creación de tests: a) delimitación conceptual del constructo objeto de evaluación; b) información sobre la construcción y evaluación cualitativa de ítems; c) resultados del análisis estadístico de los ítems; d) evidencias empíricas de la estructura interna de la prueba; e) resultados de la estimación de la fiabilidad; f) evidencias externas de la validez de la puntuaciones. Se finaliza el trabajo reflexionando sobre el alcance de las directrices propuestas y sobre la importancia de seleccionar bajo criterios claros los tests que vayan a usarse en una investigación.

PALABRAS CLAVE. Normas para la revisión de estudios instrumentales. Construcción de tests. Adaptación de tests. Selección de tests. Estudio teórico.

RESUMO. Neste trabalho discutem-se os critérios a considerar na hora de seleccionar os testes de avaliação psicológica num contexto de investigação. Tradicionalmente a atenção tem-se centrado e centra-se sobre as fases que devem orientar todo o processo de construção / adaptação de testes. Critérios standards internacionalmente aceites servem para dirigir este trabalho, e a comunidade científica coincide na importância que lhes atribui. No entanto, e para além de qualquer processo de construção/adaptação, o facto é que o uso de testes é uma constante na investigação psicológica, e uma selecção adequada das provas torna-se num assunto de grande importância. Por isso, e esquematizando em primeiro lugar os critérios que devem guiar a construção / adaptação de testes, neste estudo teórico desenvolvem-se algumas directrizes gerais a ter em consideração na altura de seleccionar testes para efectuar una investigação psicológica. A informação apresentada está organizada num total de seis pontos, cada um dos quais corresponde a uma fase distinta dentro do processo de criação de testes: a) delimitação conceptual do construto objecto de avaliação; b) informação sobre a construção e avaliação qualitativa dos itens; c) resultados da análise estatística dos itens; d) evidências empíricas da estrutura interna da prova; e) resultados da estimação da fiabilidade; f) evidências externas da validade das pontuações. O trabalho termina com reflexões sobre o alcance das directrizes propostas e sobre a importância de seleccionar sob critérios claros os testes que venham a usar-se numa investigação.

PALAVRAS CHAVE. Normas para a revisão de estudos instrumentais. Construção de testes. Adaptação de testes. Selecção de testes. Estudo teórico.

Introduction

In current psychological research, the use of instruments or tools such as computers, recording systems and measuring instruments, among others, represents a defining feature of research itself. In fact, scientific research as we know it today would be impossible without these instruments. Such instruments, as well as their influence on research findings, need to be carefully and regularly analyzed (Sturm and Ash, 2005). Among the multiple and diverse instruments that can be used in a context of psychological research, the use of assessment tests is more than frequent. Yet, we must not forget that these tests are also widely used in the professional practice generated by psychology as a discipline (Muñiz *et al.,* 2001).

The fact is that psychologists work with phenomena that cannot be directly observed and yet which they intend to measure. To do so, they use indirect approaches. To measure these phenomena it is necessary to obtain observable indicators, which highlights the importance of responses to a test as essential material for psychologists. These responses generate scores that are eventually used for multiple purposes, such as testing theories, making decisions about the effectiveness of a psychological treatment, verifying the impact of one or several independent variables experimentally, and so on. Scores obtained from tests have extremely important implications for the final result of any research using them. They also bear a great importance on the applied consequences derived from the activity of professionals who make decisions in their everyday work based on the results generated by such tests (Padilla, Gómez, Hidalgo, and Muñiz, 2006, 2007).

The Standards for Educational and Psychological Testing (AERA, APA, and NCME, 1999) are aimed at providing answers to the questions generated in the process of test creation/adaptation and use. For researchers directing their efforts to test creation/adaptation, these standards are a point of reference that guides their work and unifies assessment criteria. In fact, the debate about standards is continuously open (Linn, 2006). Besides, there are publications devoted to the optimization and improvement of standards (Koretz, 2006; Wise, 2006). This has led to the existence of guidelines that respond to particular needs, and which are undoubtedly valuable for improving the work carried out by researchers. Yet, in spite of the importance of such standards, their use is more linked to researchers who focus their efforts on so-called instrumental studies, which involve the "development of tests and devices, including both their design (or adaptation) and the study of their psychometric properties" (Montero and León, 2005, p. 124). This does not mean, however, that these standards cannot have important implications for researchers using tests for other purposes than those proper to instrumental studies.

Currently, any researcher who wishes to carry out a study requiring the use of tests usually has a choice between several possible instruments with similar objectives. In these cases, and given the direct influence of the use of a given instrument instead of another one on the final results, the informed selection of the tests must be a necessary criterion to fulfill. Choices justified by easier access to some tests rather than others or any other reasons that do not follow a scientific procedure should be avoided. Yet, the theoretical importance of the choice of instrument does not seem to be reflected in the

literature. Rather than concluding that test selection is governed by non-scientific criteria, we should say that in many cases there is a lack of information about the reasons that have led to the choice of a given test. For example, Hogan and Agnello (2004) showed that only 55% out of 696 scientific publications in which tests were used provided some evidence of the validity of the scores generated by the instruments used. Besides, as can easily be seen, a great majority of authors justify the use of the tests chosen by merely pointing out the numerical values of their reliability and validity coefficients. By doing this, they avoid taking any responsibility for the selection and use of the tests. Yet, it is known that, at the end of any research process, responsibility for the results obtained does not lie with the creators of the tests but with the authors of the research.

To make matters worse, it is a fact that most of the published tests – dealt with in journals of all kinds or produced by companies specialized in test construction and marketing – do not fulfill the minimum requirements set by the *Standards for Educational and Psychological Testing* (AERA *et al.,* 1999). Many and very different tests are developed, sometimes by specialists, and, on many occasions, by researchers far removed from this field. A great percentage of these tests is produced to satisfy very specific research needs, and use of these tests yields very little psychometric knowledge. When reviewing many of the tests published, one has to make an effort first to intuitively obtain the definition of the variable assessed. The reliability and validity coefficient values are usually found immediately afterwards and there is a lack of necessary information about the procedure followed, its justification or other aspects to consider when judging the quality of any test.

Assuming that a test satisfies the minimum scientific requirements just because it has been published is risky, to say the least. Thus, and regarding published tests, we wish to highlight that no or very little information is usually provided about the item edition process, the justification of the number of items necessary to represent the assessed construct, the proper representation of the dimensions through the items considered, and so on. What is more, there is an almost total lack of data about the controls used – both qualitative and quantitative – to guarantee the quality of the items, and the criteria used for item deletion, replacement and modification, among others.

Many problems can also be found regarding the suitability of the procedures followed to calculate reliability, or the strategies used to show the validity evidence of the test scores. As an example, let us use some of the results associated to the classic PMA (Primary Mental Abilities) test of Thurstone and Thurstone in its Spanish adaptation (TEA, 1996). When the reliability coefficient of the numerical factor is given, its value is .99. This clearly alarming finding should be a warning sign for researchers and applied psychologists who choose to use this test. The explanation to this unexpected and unacceptably high value for the reliability coefficient is the use of an inappropriate procedure: the test consists of items of speed, and was divided into two halves to calculate the coefficient mentioned (TEA, 1996, p. 13).

Regarding the validity evidence of the scores of the tests themselves, it is essential that their authors highlight and justify a syntactic definition. This definition should present with greater or lesser strength the connections of the measured construct with other constructs that make up a well-established conceptual network, or, at least, with

empirical indicators that allow making the necessary verifications later. If these conditions are not met, the validity evidence of the different tests is no more than a set of isolated statistical results that cannot be given a meaning or use other than that of hiding the shortcomings of a poor construction process.

The arguments put forward so far must seem alarming, considering the importance of using tests in psychological research. Apart from the direct influence on the results, we should also deal with the scope or widespread use of tests in most publications. For example, in a publication such as this one, the *International Journal of Clinical and Health Psychology*, 100% of the original studies published in 2007 used tests to carry out the research. Thus, it is necessary to use certain criteria to select the tests before using them, given that the fact that a test has been published does not guarantee its quality. However, in a context delimited by scientific papers, the debate should not be centered on the scientific quality of the measures used. In this field, such quality should be taken for granted as a basic need of any research. The discussion should rather be: have the tests used in published studies been selected following objective decision criteria?; have the differential aspects of various instruments constructed with similar assessment objectives been considered?; do the criteria used make it possible to conclude with greater confidence that the instrument used is the best choice among all the available ones?

The purpose of this theoretical study (Montero and León, 2005) is to propose some general guidelines for test selection in a research context, bearing in mind that many of the criteria proposed should also be taken into account by practitioners. Obviously, such a choice must be governed by the fact of being able to guarantee that the instrument chosen satisfies minimum scientific requirements, which would imply that the internationally accepted standards for test construction have been followed (AERA *et al.,* 1999). Such standards were recently discussed, and some basic guidelines were proposed for the development and review of instrumental studies (Carretero-Dios and Pérez, 2005). This study builds on such guidelines (Annex 1), highlighting that any researcher intending to use tests that are already available and therefore have been subject to prior scientific analysis must make a responsible decision. This study is set in a more general framework that deals with the standardization of the scientific procedures (Blanton and Jaccard, 2006; Botella and Gambara, 2006; Ramos-Álvarez, Valdés-Conroy, and Catena, 2006).

Criteria for test selection

The guidelines we are about to present have a specific implementation context, whichever one where it is necessary to use objective measuring instruments, in an applied or research area, and regardless of the category where such instruments can be classified: self-reports, questionnaires, psychological tests in general, and so on. Our argument is that, whenever it is necessary to assess a construct with a specific test to carry out a study, it should be done by using general guidelines so that the best instrument available is chosen and any shortcomings in these tests can be seen. However, it should be taken into account that the content of this study is influenced by the place where it

is published, and by the aim to include contents that are significant for its audience from the outset. This will be reflected in the examples and the publications used to illustrate certain issues.

For the purposes of this study, the term construct is understood as "the concept, attribute, or variable that is the target of measurement. Constructs can differ in their level of specificity from molar-level, latent variable constructs such as conscientiousness to microlevel, less inferential variables such as hitting and alcohol ingestion" (Haynes, Richard, and Kubany, 1995 p. 239). Despite this definition, it must be noted that the variables assessed in psychology are essentially constructs referring to general attributes of the individuals assessed. The definition of these constructs must be approached so that the level of specificity of the construct can be dealt with in a much more specific way. As we shall see next, this has important consequences for the selection of tests, and specifically for the stage in which the definition provided for the constructs assessed must be analyzed.

Next, we present the recommendations that should be taken into account for test selection. The recommendations are articulated into six sections, each of which corresponds to a crucial stage in the process of test construction/adaptation (for greater detail, see Carretero-Dios and Pérez, 2005, or a summary in Annex 1). Therefore, researchers should analyze these stages and check how they are reflected in the instruments they intend to use. The structure of the study is based on the assumption that the person in charge of selecting a test has considered, above all, the purpose of the assessment and what it will be used for. Thus, our exposition starts from the moment when the person involved in test selection is faced with different possible alternatives for the same assessment purpose and planned use of the scores. The following sections will guide our presentation: a) conceptual definition of the construct to assess; b) information about item construction and qualitative assessment; c) results of the statistical analysis of the items; d) empirical evidence of the internal structure of the test; e) results of the reliability estimation; and f) external evidence of score validity.

Conceptual definition of the construct to assess

It is obvious that when selecting a test, the person in charge must have a clear idea of what is to be assessed. To answer to the question of what is assessed, it is not enough to check that the name that defines the test contains a label showing its purpose, such as depression, social anxiety, sensation seeking, and so on. The most important part in constructing an instrument with adequate psychometric guarantees is to start with a complete and thorough definition of the construct assessed (Nunnally and Berstein, 1995). In fact, an ambiguous and non-specific definition leads to ambiguous and non-specific items, and therefore to scores that are not specific and whose final meaning is difficult to establish.

There are currently many tests that use the same label in their assessment purpose, which does not mean they share the same concept. The same label hides different conceptual approaches, different definitions, and therefore different – although not always explicit – measuring objectives. When deciding which test to select, the definition of the construct assessed should be consulted if available. Researchers intending to

carry out a study for which it is essential to work with specific tests have specific research objectives. Therefore, to meet their objectives they must check that the instruments they choose focus on their concept of interest beyond just sharing the same label with other instruments.

By adopting this procedure, that is, analyzing the definitions provided by the creators of tests, those in charge of the selection will realize that studies presenting a scale based on a non-specific conceptual definition of the construct assessed are more common than one might expect. The definition is usually based on a generic statement of what the construct is, which in turn is based on other constructs that are not delimited. However, this does not correspond to the recommendations issued by specialized studies (see Murphy and Davidshofer, 1994; Walsh, 1995).

The author or authors of a test should be required to specifically delimit the components or facets that define the construct to assess and operationally specify what each of these components refers to. This is known as the semantic definition of the variable (Lord and Novick, 1968). Due to the complexity of psychological constructs, a detailed and justified presentation of this definition would exceed the usual limits of a research paper. Yet, the paper should at least include a reference that makes it possible to consult the definition in detail, without limitations of space (for example, in the test manual or a book on the construct assessed, among others). Anyone in charge of selecting a test must understand that a test that does not clearly present the differentiating elements of the construct assessed, include its diverse operational expressions or clearly specify its components will lead to an imprecise construction/adaptation process, with poor content validity evidence (Downing and Haladyna, 2004; Haynes *et al.,* 1995; Smith, 2005).

When selecting a test, there would be a greater guarantee that an appropriate operational definition of the construct has been made if it was clearly shown that the authors have followed the existing recommendations about how to present this definition. More specifically, this would involve proving that they have used a table of test specifications that includes all the information of interest about the construct assessed (Osterlind, 1989). Besides verifying the existence of a detailed definition of the construct, it must be checked whether such definition was reviewed by experts before the items were created (see Carretero-Dios, Pérez, and Buela-Casal, 2006). Although it is common for this review by experts not to take place, it has been considered as an essential element to provide theoretical evidence of content validity (Rubio, Berg-Weger, Tebb, Lee, and Rauch, 2003). It also contributes to developing more representative items for the construct of interest from the first stage or stages of test construction. Thus, the table of test specifications is finally established once the review of the definition by experts has taken place (Spaan, 2006). This table should specify the construct to assess, its components and how they should be represented in the final instrument depending on their differential importance.

The existence of a table of test specifications is thought to be crucial to facilitate the adaptation of scales to different cultures (Balluerka, Gorostiaga, Alonso-Arbiol, and Aramburu, 2007). The table is an essential tool to ensure the adaptations are conceptually equivalent to the source scales. In fact, the important issue in adaptations is not just to

show evidence of a possible linguistic equivalence between the source instrument and the adapted one, which seems to be the only concern for the authors of adaptations most of the time. Instead, the key issue is to show that the adaptations are equivalent to the source test from a conceptual point of view. The existence of this table should be considered when establishing the required conceptual connection. Therefore, when selecting a test – whether the purpose is to assess original scales or their possible adaptations – it should be noted whether there is a table of test specifications or not (Spaan, 2006).

Lord and Novick (1968) also highlighted the importance of specifying the syntactic definition of the variable, that is, the relations expected between the construct assessed and other constructs or indicators once the construct has been made operational. When selecting an instrument, it must be understood that it is the network of verified relations what will finally give the scores their meaning or use. Thus, such relations must be considered as hypotheses to be checked. This eventually makes it possible to obtain evidence of the external validity of the instrument, which is an essential element of its validity as a construct (Smith, 2005).

To summarize this section, we wish to underline that the author or authors of a study using a given test must make it clear that the operational definition of the construct of interest and the way it was reached have been taken into account in the selection of the test. It must be noted as well that such definition falls within a theoretical framework of relations, which makes it possible to give a meaning to the work done with the scale.

Information about item construction and qualitative assessment

In studies presenting data on the creation/adaptation of a test, it is unusual to find information about the criteria used for item creation, a justification of the response options, and so on. There are studies available to guide this process (Martínez, Moreno, and Muñiz, 2005; Moreno, Martínez, and Muñiz, 2006; Osterlind, 1989). When choosing between instruments, preference should be given to those for which there is at least a record of the reference criteria used. This is an essential issue, given that the items are no more and no less than the specific operational expression of the components to assess. Inappropriate items always lead to a wrong operational definition, and therefore to final results far removed from the initial objectives.

The person in charge of selecting an instrument must have a clear idea of what responses are of interest regarding a construct and find out which test is best suited for this purpose. For example, in some psychological disorders, one might be interested in their frequency of occurrence. In others the goal might be to assess their intensity at the present time. In this case, depending on the purpose, one should choose a test where the items and their response format are focused on intensity or frequency.

Test creators/adaptors should be required to use the so-called table of item specifications (Osterlind, 1989; Spaan, 2006), and at least insert it in the test manual or a similar publication. This table summarizes all the issues regarding the items generated (format, response scale, proportion within the scale, examples used, and so on). This table guarantees a targeted and standardized item creation, and thus better quality items. The existence of a table of item specifications as an element that has guided item creation should be an element to consider when choosing between different instruments.

Please note the issues discussed in the previous section regarding the importance of the table of test specifications for adaptation processes, which also apply to the table of item specifications. However, it should be underlined that, in cases in which the instruments to choose from are adaptations – whose items are often translations of the originals – it must be verified that the existing recommendations on the translation process have been followed (Balluerka *et al.,* 2007; Gordon, 2004; Hambleton, 1994, 1996; Hambleton and Jong, 2003). The necessary conceptual equivalence between original and translated items should not be forgotten.

It is also necessary to check whether, once the items were created, as well as the instructions of the scale and the remaining formal aspects of the future instrument, the authors subjected them to an assessment so as to discover errors in the instructions or the wording of the items, etc. Besides, when using a test, we must have data that prove that its items are theoretically relevant for the components of the construct (Clark and Watson, 2003). Therefore, the test chosen should be examined to find out whether it provides information that guarantees that the items created are theoretically relevant for each component. The components should be represented by an appropriate proportion of items. In other words, it is important to check whether the authors of the instrument provide results about the content validity of the test (Armstrong, Cohen, Eriksen, and Cleeland, 2005; Haynes *et al.,* 1995). In this assessment process of the formal aspects of the test and the theoretical relevance of the items, certain elements are usually deleted. When selecting a test, it is important to make sure that the authors report what they deleted and why, since it provides valuable information on what remains and the strategy that was followed.

Results of the statistical analysis of the items

An essential issue to consider in the process that will finally lead to selecting a given test is related to the metric properties of the items of the instrument. Once it has been checked that the creators started with a battery clearly including more items than those needed, and that the remaining items have gone through the relevant formal and theoretical filters, one must look at the results of the statistical analysis of the items. The construction of the scale should involve a first statistical analysis of the items, like a pilot study, and the criteria for item deletion should be clearly specified. The results of this pilot study should be corroborated with a larger sample. In both cases, the sample of participants should have similar characteristics to those of the population the scale will eventually assess.

Regarding the inspection of the item statistics, the researcher must know exactly what the scale will be used for, and thus consider whether the statistics available make it possible to conclude that the test is suitable for this specific purpose. There are no universal statistic criteria that should be applied to all items regardless of what scale they belong to. Therefore, perhaps the most important issue when reviewing the analyses of items associated to an instrument is the following: to check whether the decision of deleting or keeping an item was exclusively based on a systematic use of certain numerical indices, or whether such criteria were considered taking into account the definition of the initial construct and the implementation objectives (for a review of the

most commonly used statistics and how to assess them, see Muñiz, Hidalgo, García-Cueto, Martínez, and Moreno, 2005).

Empirical evidence of the internal structure of the test

When assessing test dimensionality, the goal is to estimate "the degree to which the test items and components make up the construct to be measured and upon which the interpretations will be based" (Elosua, 2003, p. 317). Thus, conclusions about whether the internal structure of an instrument faithfully represents the components or dimensions of the construct cannot be based on the theoretical assumptions of the authors of the test or on the apparent coherence shown by the items. In order to draw this kind of conclusion, it must be checked that some procedure has been used that makes it possible to empirically assess the internal structure of the test. If the authors of the instrument start with a clear definition of the construct and its components, when inspecting the test it should be checked that a strategy has been used to test the researcher's hypothesis on how the items should be clustered.

Traditionally, and from an empirical point of view, the internal structure of tests has been explored using factor analysis (Floyd and Widaman, 1995). When selecting a test, one should be at least familiar with the details of this technique. Several studies have dealt with the inappropriate and systematic use of factor analyses or related issues (e.g. Batista-Foguet, Coenders, and Alonso, 2004; Elosua, 2005; Ferrando, 1996), especially the classic "How to fool yourself with factor analysis" (Nunnally and Bernstein, 1995 pp. 599-601). This information should be taken into account when selecting a test, and a critical assessment should be made of both the results found and the implementation process followed.

We insisted elsewhere (Carretero-Dios and Pérez, 2005) upon the idea that exploratory factor analysis knows nothing about psychology. The analysis just "clusters" similar relations. Yet, it should be stressed that the clustering may be due to more than purely conceptual elements, such as format or item type. "It should be remembered that the technique should be submitted to the conceptual interests. A cluster of items is no more than just a cluster, which may be empirically relevant but may lack any psychological meaning. There are so many "non-psychological" factors that can lead to some items being clustered with others that the application of this analysis technique in a theoretical void is totally unproductive and inefficient" (Carretero-Dios and Pérez, 2005, p. 536). Therefore, when selecting a test one should check that the application of factor analyses has been subject to theoretical premises on the dimensionality underlying the items used. Besides, the dimensionality of the test should have been verified using different samples (Elosua, 2005).

Results of the reliability estimation

The reliability of the scores of a test is another essential criterion to consider when selecting between different tests. In fact, it is one of the first subjects dealt with in specialized monographs on test construction. It is often the only value used to justify the selection of a test. In spite of this, in our presentation we have decided to deal with reliability only after discussing the aspects that come first – chronologically speaking

– in the process of constructing an assessment instrument. It is only when there is a "final" clustering of items for each component that the "final" test is available and the reliability estimation acquires its greatest scope. Yet, in many studies reliability estimations are presented in the item analysis stage, and Cronbach's alpha is usually included as an indicator among others of the item analysis. When selecting an instrument and reviewing the information available it is very important to check that the reliability estimations provided correspond to the scores obtained with the final or published version of the test, and not with earlier or experimental versions.

Again, and to assess the adequacy of a test regarding the reliability of its scores, the researcher must ask certain questions about issues closely related to the final judgment that will be issued. For example, what the scores will be used for, whether the target participants have similar characteristics to the ones used to study the test, whether the assessment conditions will be similar, and so on. Along with these questions, the concept of reliability itself and the characteristics of the different judges should not be neglected. The assessment must be reasoned and not just guided by the application of standard rules.

There are three methods commonly used to obtain reliability coefficient estimations (Traub, 1994), namely the parallel form method, the test-retest method and the single test administration method. When selecting a given instrument, an analysis should be made of the method used, the reasons for its use and its suitability for that given test. From a theoretical viewpoint, and if the test and other relevant aspects allow it (Muñiz, 1998), the best method is to repeat the test at two different moments in time. If parallel forms have been used, the researcher interested in choosing a test should deal with the problems related to this procedure, especially checking that there are really parallel forms of the test available. In any case, the test-retest method and the parallel form method face more general problems that should be considered to judge the reliability estimations provided. The most significant ones are the effect of the experience or practice of the first assessment on the second one, the "real" changes that take place in the construct assessed and the time interval used to administer the test again or to administer a parallel form of the test (Muñiz, 1998).

When reviewing published tests, it can usually be seen that the constructors/adaptors of a test tend to make reliability estimations upon a single administration of the instrument, using the procedures based on the calculation of internal consistency (Osburn, 2000). In the case of Likert scale items, the most commonly used internal consistency index is Cronbach's alpha, which is often applied without following the recommendations for its use (Cortina, 1993). Several examples have already been provided (Carretero-Dios and Pérez, 2005) to show some problems associated with the indiscriminate use of Cronbach's alpha or the superficial interpretation of the results it provides. However, the inadequate use of this index has been observed so often that we shall look at this issue in greater detail.

The person selecting a test must make sure that reliability estimations obtained through an internal consistency index are calculated for the scores of each of the components that the construct assessed is thought to include. Constructs are usually delimited by several facets or components postulated as elements that should be considered

separately. Therefore, the internal consistency should be estimated for each facet of the construct.

The judgment on the reliability obtained with Cronbach's alpha must be closely related to the format of the items or to some of their metric properties that are very linked to the final result of Cronbach's alpha. Item difficulty is an example of this. Sometimes the items used are very general questions or statements with a very similar format and common response options. This can lead the response of participants in self-reports to be "consistent" across items. However, this result may reflect a consistency across items that is more related to a factor called "instrument format" than to the theoretically assumed underlying concept. Besides, this situation could be linked to the "artificially" high values that can be obtained with Cronbach's alpha. Although researchers usually consider these high values to be a very positive finding, these data show a serious problem in the representation of the construct by the items (consult the classic problem known as the attenuation paradox, Loevinger, 1957). "In psychology, internal consistency values around .95 show a problem of under representation of the construct and inadequate validity, rather than poor reliability" (Carretero-Dios and Pérez, 2005 p. 541).

Using values that can be considered as guidelines instead of unjustified thresholds, we could say that reliability indexes around .70 are appropriate if the scale is used for research purposes. When the test is used for diagnostic or classification purposes, the minimum value advisable should be around .80 (Nunnally and Bernstein, 1995).

External evidence of score validity

External validity evidence is based on the analysis of the relations between the score or scores provided by the test and: a) a criterion that is expected to be predicted; b) other tests with the same measuring objective or with other constructs that can be expected to be related to it; and c) other variables or constructs that are expected not to be related to it, or to be less related to it than other variables (AERA *et al.,* 1999).

When we started presenting the guidelines for selecting an assessment test we underlined the following ideas: the target construct should be operationally (semantically) defined; there should also be a conceptual (syntactic) definition delimited by the relations expected with other constructs, that is, the construct should be located in a network of theoretical relations. The task of the person intending to select a test would be to establish to what extent the test scores have produced evidence that confirms the expected relations. The inspection of these findings provides the researcher with the information about the usefulness or meaning of the test scores.

Those interested in selecting a given test should bear in mind that there is no methodological strategy or statistical analysis technique that is exclusive of studies aimed at obtaining external evidence of validity. The results could be derived from using experimental, quasi-experimental or non-experimental strategies, so the analysis techniques could be seen as diverse. Because of this, what is really important in this context is to realize whether the authors of a given test have justified the relations presented on the basis of the theories of interest or results or earlier research. This should be reflected in the syntactic definition of the variable. Of course, it must be

checked whether, depending on the specific objectives of analysis, the study methodology and the analysis procedures used are the most suitable ones. This is applicable to the scientific review of any study published. Besides, it must not be forgotten that the scores of a test do not produce evidence that establishes its validity once and for all. By definition, obtaining validity evidence implies an unfinished process that is subject to permanent review and is sensitive to the evolution of knowledge on the measured construct. Those responsible for test selection should be sensitive to these aspects as well.

Conclusions

The choice of a test to be used in research is an issue of great importance. This study was carried out to discuss the possible difficulties that can be found in this process of test selection and provide some guidelines to facilitate this choice. However, the purpose of the guidelines proposed is not to become a simplified and specific implementation guide. Instead, they should be understood as a tool that leads the user to think about certain elements and consider some decisions more cautiously. A researcher will never be able to draw rigorous conclusions if the raw material used to produce them is scores provided by inadequate instruments. Likewise, and given the ethics that define scientific work, the person in charge of a study should not be satisfied with the fact of using a test with some psychometric support and with sufficient scientific guarantees. Instead, there should be basic information that guarantees that the choice made is the best possible one among all the alternatives whose existence is known to this person.

A research report in a regular scientific journal has restricted space. The justification of why a given instrument has been chosen would exceed this space. Therefore, it would not be feasible to provide a comprehensive account of the reasons that have led the authors to use a given test instead of any of the possible alternatives. However, this does not prevent the author or authors of a research study from using the guidelines presented here or any other method that guarantees a scientific selection of the tests. There are other sections of a report in which some information is left out in order to simply mention the procedure followed or the strategy used. Likewise, in this field of test selection editors and reviewers of scientific publications should insist that that the authors of a study mention at least the criteria followed to select the instruments and say where these criteria are dealt with in greater detail. It is surprising to find that the instruments section in many scientific journals includes only a list of scales, for which the only information available is, at most, their reliability and some references where they have been applied to be studied. Next to the indication of the scales used, a question should serve as a theme for this section: Why these tests instead of others? The authors of any scientific study using psychological assessment tests should be able to answer this question.

References

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

Armstrong, T.S., Cohen, M.Z., Eriksen, L., and Cleeland, C. (2005). Content validity of self-report measurement instruments: An illustration from the development of the Brain Tumor Module of the M.D. Anderson Symptom Inventory. *Oncology Nursing Forum, 32,* 669-676.

Batista-Foguet, J.M., Coenders, G., and Alonso, J. (2004). Análisis factorial confirmatorio. Su utilidad en la validación de cuestionarios relacionados con la salud. *Medicina Clínica, 122,* 21-27.

Balluerka, N. Gorostiaga, A., Alonso-Arbiol, I., and Aranburu, M. (2007). La adaptación de instrumentos de medida de unas culturas a otras: una perspectiva práctica. *Psicothema, 19,* 124-133.

Blanton, H. and Jaccard, J. (2006). Arbitrary metrics in Psychology. *American Psychologist, 61,* 27-41.

Botella, J. and Gambara, H. (2006). Doing and reporting a meta-analysis. *International Journal of Clinical and Health Psychology, 6,* 425-440.

Carretero-Dios, H. and Pérez, C. (2005). Normas para el desarrollo y revisión de estudios instrumentales. *International Journal of Clinical and Health Psychology, 5,* 521-551.

Carretero-Dios, H., Pérez, C., and Buela-Casal, G. (2006). Dimensiones de la apreciación del humor. *Psicothema, 18,* 465-470.

Clark, L.A. and Watson, D. (2003). Constructing validity: Basic issues in objective scale development. En A.E. Kazdin (Ed.), *Methodological issues & strategies in clinical research (3rd ed.)* (pp. 207-231). Washington, D.C.: APA.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Downing, S.M. and Haladyna, T.M. (2004). Validity tretas: Ivercoming interferente with proponed interpretations of assessment data. *Medical Education, 38,* 327-333.

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

Elosua, P. (2005). Evaluación progresiva de la invarianza factorial entre las versiones original y adaptada de una escala de autoconcepto. *Psicothema, 17,* 356-362.

Ferrando, P.J. (1996). Evaluación de la unidimensionalidad de los ítems mediante análisis factorial. *Psicothema, 8,* 397-410.

Floyd, F.J., and Widaman, K.F. (1995). Factor análisis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286-299.

Gordon, J. (2004). Developing and improving assessment instruments. *Assessment in Education: Principles, Policy and Practice, 11,* 243-245.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229-240.

Hambleton, R.K. (1996). Adaptación de tests para su uso en diferentes idiomas y culturas: fuentes de error, posibles soluciones y directrices prácticas. En J. Muñiz (Ed.), *Psicometría* (pp. 203-238). Madrid: Universitas.

Hambleton, R.K. and Jong, J.H. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20,* 127-134.

Haynes, S.N., Richard, D.C.S., and Kubany, E.S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238-247.

Hogan, T.P. and Agnello, J. (2004). An empirical study of reporting practices concerning measurement. *Educational and Psychological Measurement, 64,* 802-812.

Koretz, D. (2006). Steps toward more effective implementation of the Standards for Educational and Psychological Testing. *Educational Measurement: Issues & Practice, 25,* 46-50.

Linn, R.L. (2006). Following the Standards: Is it time for another revisions? *Educational Measurement: Issues & Practice, 25,* 54-56.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694.

Lord, F.M. and Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Martínez, R.L., Moreno, R., and Muñiz, J. (2005). Construcción de ítems. En J. Muñiz, A.M. Hidalgo, E. García-Cueto, R. Martínez, and R. Moreno (Eds.), *Análisis de ítems* (pp. 9-52). Madrid: La Muralla.

Montero, I. and León, O. (2005). Sistema de clasificación del método en los informes de investigación en Psicología. *International Journal of Clinical and Health Psychology, 5,* 115-127.

Moreno, R., Martínez, R.J., and Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2,* 65-72.

Muñiz, J. (1998). *Teoría clásica de los tests.* Madrid: Pirámide.

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

Muñiz, J., Hidalgo, A.M., García-Cueto, E., Martínez, R., Moreno, R. (2005) *Análisis de ítems*. Madrid: La Muralla.

Murphy, K.R. and Davidshofer, C.O. (1994). *Psychological testing: Principles and applications* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Nunnally, J.C. and Bernstein, I.J. (1995). *Teoría psicométrica.* Madrid: McGraw-Hill.

Osburn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5,* 343-355.

Osterlind, S.J. (1989). *Constructing Test Items.* London: Kluwer Academic Publishers.

Padilla, J.L., Gómez, J., Hidalgo, M.D., and Muñiz, J. (2006). La evaluación de las consecuencias del uso de los tests en la teoría de la validez. *Psicothema, 18,* 307-312.

Padilla, J.L., Gómez, J., Hidalgo, M.D., and Muñiz, J. (2007). Esquema conceptual y procedimientos para analizar la validez de las consecuencias del uso de los tests. *Psicothema, 19,* 173-178.

Ramos-Álvarez, M.M., Valdés-Conroy, B., and Catena, A. (2006). Criterios para el proceso de revisión de cara a la publicación de investigaciones experimentales y cuasi-experimentales en Psicología. *International Journal of Clinical and Health Psychology, 6,* 773-787.

Rubio, D.M., Berg-Weger, M., Tebb, S.S., Lee, E.S., and Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27,* 94-104.

Smith, G.T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment, 17,* 396-408.

Spaan, M. (2006). Test and item specifications development. *Language Assessment Quarterly, 3,* 71-79.

Sturm, T. and Ash, M.G. (2005). Roles of instruments in psychological research. *History of Psychology, 8,* 3-34.

TEA (1996). *PMA: Aptitudes Mentales Primarias* (9ª edición). Madrid: TEA.

Traub, R.E. (1994). *Reliability for the social sciences: Theory and applications.* London: Sage.
Walsh, W.B. (1995). *Tests and assessment.* New York: Prentice-Hall.
Wise, L.L. (2006). Encouraging and supporting compliance with Standards for Educational Tests. *Educational Measurement: Issues & Practice, 25,* 51-53.

ANNEX 1. Standards for the development and review of instrumental studies (from Carretero-Dios and Pérez, 2005).

(A) Justification of the study

		Yes	Not sure	No
A.1.	The justification is based on relevant background.			
A.2.	The creation/adaptation of the instrument will be a significant contribution to the area of study.			
A.3.	The general assessment purpose of the test is clearly specified.			
A.4.	The target population of the test is specified.			
A.5.	The aim or purpose that the test scores will be used for is specified.			
A.6.	The creation/adaptation process is viable.			

(B) Conceptual definition of the construct to assess

		Yes	Not sure	No
B.1.	The most relevant attempts to conceptualize the construct of interest are clearly specified.			
B.2.	The different conceptual proposals presented are integrated into one or several theoretical frameworks of reference.			
B.3.	A review is made of the main assessment instruments aimed at assessing this construct or related constructs.			
B.4.	After the review there is an operational proposal of the facets or operational components of the construct to assess that is judged by experts.			
B.5.	The information related to judgment by experts (selection of experts, material used, assessment method, etc.) is presented in detail.			
B.6.	The operational definition of the construct is finally established considering the results of the assessment by experts, the research data and the theoretical frameworks of reference.			
B.7.	The relations expected between the construct and other variables are specified considering the definition adopted for the construct.			
B.8.	The relations predicted for the total scores in the construct are well justified.			
B.9.	If the construct includes different facets or components, the relations expected for each of these components are also specified.			
B.10.	The relations predicted are clearly presented, specifying when the construct will be a predictive valuable, a predicted variable or a covariate.			

(C) Item construction and qualitative assessment

		Yes	Not sure	No
C.1.	The information justifying the type of items to construct (including the format, type of wording, response scale, etc) is clearly presented.			
C.2.	The author uses a table of item specifications to guide the development of the items.			
C.3.	The table of item specifications includes all the information necessary for the construction of the items.			
C.4.	The final number of items of the scale that is to be created/adapted is well justified.			
C.5.	The initial battery of items is formed by at least twice as many items by component than those planned to be finally used.			
C.6.	If the items have been translated, a strategy ensuring the conceptual equivalence between the original and the translated items has been used.			
C.7.	If the items have been translated, the author provides new items related to the components of the construct to assess.			

C.8.	The content validity evidence provided by the assessment of the initial battery of items by a group of judges is presented.			
C.9.	All the information related to the procedure used by a group of judges to assess the items is presented.			
C.10.	The assessment of the items by a group of judges has been properly carried out.			
C.11.	The items deleted after completion of the assessment by a group of judges are clearly specified.			
C.12.	The items kept after completion of the assessment by a group of judges are clearly specified.			

(D) Statistical analysis of the items

		Yes	Not sure	No
D.1.	The study is clearly defined (first study of the items, pilot study or cross-validation).			
D.2.	The objectives of the analysis are clearly specified (homogeneity and consistency of the scale *versus* criterion validity).			
D.3.	All the information is provided regarding the items, instructions to participants, application context, etc.			
D.4.	The study sample has similar characteristics to those of the target population of the test.			
D.5.	The sample size is adequate for the objectives of the study.			
D.6.	The sampling procedure is similar to the one planned for the final scale.			
D.7.	The criteria to consider for item selection-deletion are clearly specified.			
D.8.	The statistical calculations made are relevant.			
D.9.	The results – both qualitative and quantitative – are clearly discussed.			
D.10.	Theoretical issues are taken into account in decisions about items.			
D.11.	It is clearly specified which items are deleted and why.			
D.12.	The items selected are clearly specified.			

(E) Study of the dimensionality of the instrument (internal structure)

		Yes	Not sure	No
E.1.	The study is clearly defined (first study of the dimensionality of the scale or cross-validation of earlier results).			
E.2.	The objectives of the analysis are clearly specified (exploratory study *versus* confirmatory study, or both).			
E.3.	The information presented clearly justifies the objectives proposed.			
E.4.	All the necessary information is provided to inform the readers about the background justifying the scale and the dimensionality expected for the scale.			
E.5.	The information about the sample is complete and relevant.			
E.6.	The study sample has similar characteristics to those of the target population of the test.			
E.7.	The sample size is adequate for the objectives of the study.			
E.8.	The sampling procedure used is adequate for the objectives of the study.			
E.9.	If an exploratory factor analysis procedure is used, the need for it is justified.			
E.10.	The reason why a specific type of exploratory factor analysis has been chosen instead of another one is clearly explained.			
E.11.	Before applying the exploratory factor analysis, the author informs about the adequacy of the correlation matrix (Bartlett's test of sphericity and the Kaiser-Meyer-Olkin measure).			
E.12.	The dimensionality of the scale is interpreted on the basis of the			

	rotated factor solution.			
E.13.	The factor rotation procedure used is well justified.			
E.14.	The factor rotation procedure used is adequate.			
E.15.	The information provided about the resulting factor solution is adequate (number of factors, relevant factor loadings of the items they contain, percentage of variance explained and commonality).			
E.16.	The statistical procedures used to discuss which factors are relevant and should be considered are adequate.			
E.17.	The discussion about the factors to consider is set in the framework of earlier theoretical and empirical research.			
E.18.	If a procedure based on a confirmatory factor analysis is used, the measuring model (that is, the way of distributing the items) to analyze is clearly established.			
E.19.	The study makes a comparative diagnosis of alternative proposals besides the reference model.			
E.20.	The estimation procedure used is justified.			
E.21.	The estimation procedure chosen in the study is adequate.			
E.22.	The author uses several indices simultaneously for the diagnosis of the model.			
E.23.	The study informs of the reasons for choosing certain indices and of the threshold values that should be considered to assess the goodness-of-fit of the model.			
E.24.	The results for the different goodness-of-fit indices are clearly presented in the study.			
E.25.	If the author makes changes to improve the fit, the decisions are clearly supported (theoretically and empirically) and are clearly shown in the study.			
E.26.	The author presents the diagram (path diagram) showing the distribution of the items for each factor, the "degree" to which each of these items is predicted by the factor it corresponds to, and, more generally, all the parameters considered to be relevant in the initial specification of the model.			

(F) Reliability estimation

		Yes	Not sure	No
F.1.	The study justifies the reliability estimation procedure used (theoretical adequacy).			
F.2.	The reliability estimation method used is considered to be adequate.			
F.3.	If the *test-retest* method is used in the study, the most significant aspects affecting this calculation besides theoretical questions are provided and discussed (time interval, assessment conditions, sample correspondence, etc.).			
F.4.	The use of the *test-retest* method is appropriate considering the most significant aspects affecting its application (time interval, assessment conditions, sample correspondence, etc.).			
F.5.	If the *parallel form* method is used in the study, the most significant aspects affecting this calculation besides theoretical questions are provided and discussed (data on the equivalence of tests, as well as the information common to the test-retest method, such as time interval, assessment conditions, sample correspondence, etc.).			
F.6.	The use of the *parallel form* method is appropriate considering the most significant aspects affecting its application (equivalence of tests, time interval, assessment conditions, sample correspondence, etc.).			
F.7.	If *Cronbach's alpha* index based on internal consistency is used in the study, the most significant aspects affecting this calculation besides theoretical questions are provided and discussed (number of items for each component of the construct and item format).			

F.8.	The use of *Cronbach's alpha* is appropriate considering the most significant aspects affecting its application (number of items for each component of the construct and item format).			
F.9.	If the *split-half* method is used to calculate internal consistency, the most significant aspects affecting this calculation besides theoretical questions are provided and discussed (procedure used to obtain the two halves of the test and number of items forming them).			
F.10.	The use of the *split-half* method is appropriate considering the most significant aspects affecting its application.			
F.11.	The size of the study sample is adequate for the objectives of the research.			
F.12.	The characteristics of the participants are adequate considering the objectives of the test and the purpose of the scores.			
F.13.	The assessment procedure used is adequate considering the characteristics of the test.			
F.14.	The results derived from the reliability estimation are clearly shown.			
F.15.	The results are discussed taking into account both methodological and theoretical aspects.			
F.16.	If poor reliability data are obtained, the strategies to adopt are discussed in the study.			