Papeles del Psicólogo, 2015. Vol. 36(3), pp. 161-173
http://www.papelesdelpsicologo.es

NEW GUIDELINES FOR TEST USE: RESEARCH, QUALITY CONTROL AND SECURITY OF TESTS

José Muñiz[1], Ana Hernández[2] and Vicente Ponsoda[3]
[1]*Universidad de Oviedo.* [2]*Universidad de Valencia.* [3]*Universidad Autónoma de Madrid*

***Antecedentes**. Para llevar a cabo una evaluación psicológica rigurosa es necesario que los profesionales que la realizan tengan una preparación adecuada, que los tests utilizados muestren unas buenas propiedades psicométricas, y que se utilicen de forma correcta. El objetivo de este trabajo es presentar las directrices recientes de la Comisión Internacional de Tests sobre el uso de los tests en tres ámbitos: investigación, control de calidad y seguridad en el manejo de las pruebas. **Método**. Se revisarán y comentarán los directrices recientes desarrolladas por la Comisión Internacional de Tests. **Resultados**. Las nuevas directrices sobre el uso de los tests ofrecen todo un conjunto de recomendaciones teórico-prácticas para guiar la utilización adecuada de los tests en contextos de investigación, para desarrollar e implementar procesos de control de calidad efectivos, y para salvaguardar la seguridad de todos los datos implicados en un proceso evaluativo. **Conclusiones**. Las nuevas directrices desarrolladas por la Comisión Internacional de Tests contribuirán a una adecuada utilización de los tests en contextos de investigación, a una mejora en los procesos de control de calidad de las pruebas, y a garantizar la seguridad en los procesos evaluativos.*
***Palabras clave:** Uso de tests, Comisión Internacional de Tests, Directrices, Investigación, Seguridad, Control de calidad*

***Background**. In order to carry out a rigorous psychological evaluation, three conditions must be met: the practitioners must have the appropriate qualifications, the tests must show good psychometric properties, and they must be used correctly. The aim of this paper is to present the recent guidelines developed by the International Test Commission on the use of tests in three areas: research, quality control, and security. **Method**. The guidelines developed by the International Test Commission will be analysed and discussed. **Results**. The new guidelines on the use of tests offer a whole range of theoretical and practical recommendations to guide the appropriate use of tests in research settings, in order to develop and implement effective quality control strategies, and to preserve the security of all of the data involved in the assessment process. **Conclusions**. The new guidelines developed by the International Test Commission will contribute to the correct use of tests in research settings, to an improvement in the quality control of testing, and to ensuring security in assessment processes.*
***Key words:** Test use, International Test Commission, Guidelines, Research, Quality Control, Security.*

Tests are fundamental tools in the professional practice of psychologists and, as with any other tool, we must use them properly. The usefulness of tests is based on three fundamental pillars: the practitioners must be suitably qualified, the tests must have appropriate psychometric properties, and they must be used correctly. If these three conditions are met, the tests will be of great use to psychologists when practising their profession. The universities and other educational institutions focus their efforts on the training of practitioners, the editors aim to provide the market with the best possible tests, and the various national and international organisations strive to improve the use made of the measuring instruments. Out of all of these organisations, the European Federation of Psychologists' Associations (EFPA) and the International Test Commission (ITC) deserve special mention. At the national level, the Spanish Psychological Association (COP) works closely with the EFPA and the ITC and is a member of both organisations. These national and international organisations carry out varied activities and projects that revolve around two major strategies, which we can call *restrictive* and *informative*. Below we offer a brief description of the two strategies, following the previous studies by Muñiz and Bartram (2007), Muñiz and Fernández-Hermida (2010), and Muñiz (2012).

The *restrictive* strategy brings together a set of actions that have been implemented in order to restrict the use of the tests to those practitioners who are actually qualified to use them. The systems vary from one country to another (Bartram, 1996; Bartram & Coyne, 1998; Muñiz, Prieto, Almeida & Bartram, 1999), although one of the most common systems in several countries, including Spain,

Correspondence: José Muñiz. *Facultad de Psicología. Universidad de Oviedo. España. E-mail: jmuniz@uniovi.es*

involves classifying the tests into three categories (A, B, C) from lower to higher specialisation according to the APA criteria, with the tests in categories B (collective tests of a cognitional nature and personality tests) and C (individual scales and projective tests) being for use exclusively by psychologists. Another option is for the practitioners to obtain a specific certification which verifies conclusively that they know the tests adequately. While these restrictions and others are recommended, they alone do not guarantee the correct use of the tests (Moreland, Eyde, Robertson, Primoff & Most, 1995; Simner, 1996), it being necessary to complement this strategy with the dissemination of information to all parties involved, such as practitioners, users, institutions, and society in general.

The actions that have been carried out within the framework of the strategy that we have called *informative* refer to all kinds of initiatives aimed at disseminating information on the practical aspects of tests. It is understood that the more information that the practitioners, users, families, and in general all parties involved possess regarding the use of the tests, the lower the likelihood of the tests being misused. In this regard, the different national and international organisations have developed ethical and deontological codes as well as various guidelines for the appropriate use of tests. The most noteworthy of these are the meta-code of ethics of the EFPA (2005), the code developed by the North American Joint Committee on Testing Practices (2002) and the guidelines by the European Association of Psychological Assessment (Fernández-Ballesteros et al., 2001). See also the good reviews by authors such as Koocher and Keith-Spiegel (2007), Lindsay, Koene, Ovreeide and Lang (2008), or Leach and Oakland (2007), and particularly the last special issue devoted to this subject in the journal *Papeles del Psicólogo* (2009). As well as these codes, there is a set of guidelines currently available that marks the steps to be taken from the very construction of the test, its implementation, interpretation and the application of its results (Bartram, 1998; Brennan, 2006; Downing & Haladyna, 2006; Muñiz, 1997). Deserving special mention are the technical standards developed by the American Psychological Association together with two other organisations (AERA, APA and NCME, 2014) as well as the guidelines developed by the International Tests Commission (ITC) for the translation and adaptation of tests across cultures (Hambleton, Merenda, & Spielberger, 2005; Muñiz, Elosua & Hambleton, 2013). For other guidelines on the use of tests in general, computerised and Internet tests, or the use of tests in the workplace and organisations, see, for example, the article by Muñiz and Bartram (2007) or the ITC website (www.intestcom.org) or that of the EFPA (www.efpa.eu). Information of interest can also be found on the website of the Spanish Psychological Association, in the Test Commission section (www.cop.es). As well as the ethical codes and guidelines, there are two measures that deserve attention within the actions pertaining to the informative strategy; firstly, the new ISO-10667 standard which regulates everything concerning the evaluation of people in work contexts and, secondly, models of test evaluation developed in various countries, including Spain (Prieto & Muñiz, 2000), which are applied to the tests available in the market in order to provide information regarding their strengths and areas for improvement (Hernández, Tomás, Ferrreres & Lloret, 2015; Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez, & Peña-Suárez, 2011; Ponsoda & Hontangas, 2013).

Within this *informative* strategy, the ITC has recently developed three documents aimed at improving the use of tests in three different areas. The first is a statement on the use of tests in research, the second is a set of guidelines on the quality control of tests, and the third is a set of guidelines on the security of tests. These three documents have recently been translated into Spanish and the central objective of this paper is to present and highlight their most important contributions. The full versions in Spanish of these three documents are available on the website of the Test Commission of the Spanish Psychological Association (www.cop.es), in the Test Commission section.

THE USE OF TESTS AND OTHER ASSESSMENT TOOLS IN RESEARCH

This document was initiated by Professor Fanny Cheung and prepared for the ITC by Professors Dragos Iliescu and Dave Bartram, and was translated into Spanish by José Muñiz.

This statement by the ITC seeks to clarify all matters relating to the use of tests for research purposes. It is divided into seven parts: Permission to use the tests in research, Permission to reprint, Modifying the test or its components, Ethical use of the tests, Documentation, Conflict of Interests, and Using research tests in professional practice. A number of aspects from each section are discussed below.

PERMISSION TO USE THE TESTS IN RESEARCH

In this section, the key issue that the ITC's statement emphasises is that if the tests are copyrighted the researcher has to ask permission from the owner of the rights. Only freely accessible tests can be used without requesting permission, however, this kind of situation is rare, although some researchers are not aware of it. A very common mistake is to assume that if a test has been published in a journal or another document it must be freely accessible. The declaration of the ITC recommends that, when there is the slightest doubt, permission must be requested from the authors to use a test. Please note the words in the statement of the ITC: *Research versions of the instruments are often published in magazines or on the authors' websites. When these are published in means that are open for the public to access, it may seem as though the tests are for public use, however, by default, the authors are the holders of the copyright until they give the rights to another entity, or they give explicit permission for the free use of their tests.*

PERMISSION TO REPRINT

Here the message regarding the reprinting of test items is very clear; it cannot be done without permission: *Copyrighted studies must not be reproduced, distributed or publicly presented; nor it is permissible to carry out work derived from these studies without the permission of the owner of the copyright.* When presenting the results, the researchers may be forced to present some items in order to better describe their data, in which case either permission is requested of the holders of the copyright, or alternatively the researchers can develop items that are similar to those in the original test but without reproducing them, so that copyright is not violated.

MODIFYING THE TEST OR ITS COMPONENTS

The message of the statement by the ITC on the modification of a test is very clear: *When using tests that are subject to copyright, researchers must not modify any part of the test, as this would jeopardise the integrity of the test, violate the copyright and be considered unlawful, unless the researchers are duly authorised to do so.* Of course, if the test is for public use or the necessary permission has been obtained, modifications can be carried out in order to align the test to the conditions of use; *for example, for use in a different culture from the one in which it was generated it may require translation into another language, the removal of some aspects, the rephrasing of others, the addition of items, or modifications to the instructions or the scale of the items.*

ETHICAL USE OF TESTS

The contents of the declaration of the ITC are in line with the provisions of other standards and ethical codes regarding the use of tests in the professional sphere: *The ethical use of tests in research and in professional contexts is very similar. Those who use tests in research must act ethically and professionally, be proficient in the use of the tests, be responsible for their use and ensure the security of the materials used and the confidentiality of the results... The responsibility of a qualified professional comprises the whole assessment process, including the data collection, encoding, analysis, reporting and the application of the data in its various forms.*

DOCUMENTATION

Research requires rigorous documentation to permit replication so, when tests are used, full information must be provided regarding their characteristics and psychometric properties, with the case of newly created tests deserving special mention, as established in the ITC's statement: *Researchers building a new measuring instrument should at least provide information on the theoretical foundation of the test and its purpose, the system used for the initial selection of the items, how they were analysed subsequently and the selection criteria, the number of items in each facet, the scaling methods used, as well as information about the evidence of validity and the accuracy of the measurements, for example the reliability or other indicators of the accuracy of measurements, depending on the test measurement model.*

CONFLICT OF INTERESTS

Researchers must state their sources of funding, in case this could affect their impartiality in the research process, as the statement by the ITC indicates: *Research funded by external agencies may have a particular interest in the results, for example a government department that is trying to implement a policy, or a test publishing company that is the copyright holder.*

USING THE RESEARCH TESTS IN PROFESSIONAL PRACTICE

When the tests are used in professional, clinical, educational, work, or other contexts, their characteristics

and psychometric properties have to be very demanding, since they will be used to make important decisions affecting people's lives. These conditions can be more relaxed when the tests are used for research alone, where the data are often used in aggregate form, and the tests themselves are even being built and tested. However, we must be very cautious and we must warn of the limitations when the research versions of the tests are used in the professional field. As the ITC's statement clearly indicates: *The author of the test has a responsibility not to contribute to the use of research measuring instruments in professional practice before sufficient information has been published regarding its psychometric properties.* We trust that these recommendations from the ITC will help researchers to improve the use of tests, which in turn will contribute to improving the quality of the **G**research, which will undoubtedly result in the creation of better quality measuring instruments for professional practice.

GUIDELINES FOR QUALITY CONTROL OF THE TEST SCORES, THEIR ANALYSIS AND THE REPORTS ON THE SCORES

These guidelines were prepared for the ITC Council by Avi Allalouf and translated into Spanish by Ana Hernández. They are motivated by the fact that, as we all know, to err is human. And even though the assessment of people using tests (understood in the broadest sense of the term) is typically performed by professional experts, this type of assessment is no exception and may be subject to errors.

The errors that occur may include, among other things, the application of an erroneous scoring template, the incorrect conversion of raw scores on transformed scales, the incorrect interpretation of a score (which could depend on its prior transformation), reports being sent to the wrong client, or an excessive delay occurring in reporting the results. These errors could have significant consequences for the people being tested, for society and for the profession. For example, applying the wrong template or the incorrect transformation of a score could prevent a qualified candidate from accessing a specific job, or it could result in inadequate educational intervention, or in people being awarded academic credentials without the required knowledge and skills. Or for example a delay in reporting the results could cause problems for people who, due not to receiving the information in time, would not be able to access a particular place or institution. However errors can also affect the tests administered and their reputation, with a reduction in the reliability and validity of the scores obtained. All of this would contribute to a loss of confidence in educational and psychological tests and to the evaluation processes being questioned.

Therefore, professional users of tests must be able to anticipate any potential errors, to prevent and address them, and this is the main objective of the guidelines. Specifically, the quality control guidelines are especially focused on the mistakes that may be committed during the assignment of scores, test analysis and reporting phases, but it may also be useful to consider these guidelines for the initial phases of the evaluation process: the test construction or selection, and its administration. The guidelines are intended for situations of large-scale evaluation (educational or work), when the test is principally a measure of performance or ability (as opposed to attitudes, preferences, etc.) Thus, the guidelines are primarily aimed at the professionals involved in this type of evaluation: test constructors, administrators, editors, psychometricians, people involved in preserving the security of the test, psychologists, educators, and computer programmers, among others. However, many of the proposals included in the guidelines could also be applied to smaller-scale assessments, for different purposes, or implemented using other types of tests (interviews, work samples, etc.).

The guidelines are structured in two parts. In the first part, a series of general guiding principles are presented. In the second part, the detailed guidelines are presented step by step.

GENERAL PRINCIPLES

These deal with seven main issues:

1) The need to check if there are adequate quality control processes to apply to a specific evaluative situation, so that, if these processes do not exist or are not adequate, they can be developed, adapted, improved and implemented.
2) The need to develop agreements on the basic principles of the evaluation process among the various professionals involved and sometimes among the different parties involved. This point includes issues such as establishing the assessment objectives, allocating responsibilities, proposing a timeline, selecting the most appropriate ways of scoring, the best way to transfer the data, the type of report to make and

the recipients of such information, among others.

3) The need to ensure the availability of the necessary resources (space, financial, material, time and personnel) in order to carry out all of the phases of evaluation as planned, as well as to provide additional resources that may be needed if there should be a setback.
4) The need, at times, to make adjustments between the needs and expectations that the interested parties have in the evaluation results (the people being tested, the teachers, the parents, etc.), and the needs and expectations of those responsible for the evaluation. It is therefore advisable to have good communication between the parties. This point includes aspects such as the establishment of agreements, the responsibility for decision-making, and the possibility that the person being examined may question or review the results and make suggestions.
5) The need to ensure that the right people and the right work environment are available in order to carry out the assessment process. Here we consider the timetables and the way of working as well as the support for the staff involved in the evaluation (e.g. through training).
6) The need to have one or more independent supervisors that track the established processes of quality control and record and report any errors that may be observed.
7) The need for all those involved in the evaluation process to follow the agreed procedures regarding the documentation of activities and the recording of errors using standardised forms. Agreements should be established regarding which staff members are responsible for each stage of the process and, when mistakes are observed, they should be reported promptly to prevent the error from occurring again in the future.

DETAILED STEP-BY-STEP GUIDELINES

These are presented in five sections: the planning and design of the report, the consideration of background and biographical data, the test scores, the analysis, and preparing the report. In all of these sections a series of actions and steps are suggested to ensure the quality of the process. It is recommended that these actions are carried out explicitly with large-scale assessments. However, with smaller scale evaluations, although the principles of the guidelines are still relevant, some of the phases could be omitted or simplified. The reason is that some of the proposed procedures require significant resources and are based on models that require large samples, so they could be adapted in order to apply them to smaller samples. Some of the recommendations contained in the different sections are summarised below.

1) **The planning and design of the report**. Since all of the phases must be aimed at ensuring the quality of the product of the assessment, which is the report, it is recommended that, prior to starting the process, decisions are made regarding what to report on, what type of scores to use, in how much detail, to whom, when, etc.
2) **Background and biographical data.** Recommendations are proposed relating to this kind of information, in order to verify the identity of the people tested, whilst maintaining the confidentiality of the data, and in order to explain results that are unexpected or inconsistent with previous studies.
3) **Scores.** This section contains several subsections: a) *The collection and storage of the responses* of the people being evaluated, including recommendations on the storage of the answer sheets and electronic data, the use of identification codes, data security, and the need to ensure the correctness of the algorithms, conversion tables and scales. b) *The obtaining of scores* which includes, among other things, data analysis to ensure that the scores are within the expected range, the identification and review of extreme scores, the identification of people with excessive differences in the scores obtained on correlated sub-tests and the analysis of the psychometric properties of the items, which will facilitate the identification of errors in the correction template, omissions, etc. And finally, c) for *tests with open-response items* for performance qualification, work samples, role play, interviews, etc., which are less objective than multiple-choice tests, a number of additional precautions are mentioned. The issues included relate to conducting training courses for assigning scores, producing scoring instructions with examples, or the number of evaluators.
4) **Analysis of the test.** Here the guidelines are grouped into four subsections: a) First the *analysis of the items* is re-emphasised, prior to obtaining the total score, in order to assess their quality. b) Equivalence/calibration of new forms of the test and items, if the test is performed at various times and/or in various forms. On this point, the following recommendations, inter

alia, are made: to develop routines to ensure that the specified procedures and equivalence designs have been carried out correctly and that the assumptions on which they are based are fulfilled, to check whether different procedures based on different assumptions give similar results, to compare the scores obtained with those that were anticipated in terms of the background and biographical data collected, or, if there are cut-off points to differentiate between those being evaluated according to their level, to check the similarity of the reasons for the passes and fails in the different groups assessed. c) *The calculation of standardised scores*. On this point it is recommended, among other things, to check the adequacy and accuracy of the conversion that has been performed, manually converting some of the scores and comparing the results with those generated by the computer, to compare the results obtained with different programs, or to test the relationship between the raw scores and standardised ones using scatter plots. And finally, d) *checking the security of the tests*. This section summarises some of the main recommendations of the ITC guidelines on the security of tests, about which this article is also concerned.

5) **Preparation of reports.** There are three different subsections here: a) *Report on the scores*. This includes recommendations on the use of focus groups that enable the production of interpretative guidelines, specifying the degree to which scores can be interpreted reliably, the use of data repositories to report the results promptly, or advice from public relations experts when the reports are to be submitted to politicians and the media. b) Measures to ensure the *security of the reports*. This section provides a number of recommendations on the rectification of reports or the prevention of forgery. And finally, c) *documentation*. It is recommended that an internal report be carried out with exhaustive information on the process of obtaining scores and the key statistics obtained, in order to guarantee the accuracy of the entire process. The possibility of making public a number of statistics on group results (for example, by year, or by sex) is also discussed, providing a brief explanation on the interpretation of those statistics.

In short, the guidelines bring together a number of suggestions for quality control throughout the different stages of the assessment process. The extent to which these recommendations are being followed (or not) in a given assessment situation can always be checked before moving to a later stage, in order to verify compliance or perform corrective actions. Although some of the guidelines only apply to large-scale assessments, many of them are applicable to any assessment situation. And although it is likely that many of the professionals involved in evaluating people (psychologists and educators) follow many of the recommendations made, the systematic adoption of the ITC guidelines for quality control will help prevent mistakes that can be made when conducting assessments using tests.

GUIDELINES ON THE SECURITY OF TESTS, EXAMS AND OTHER ASSESSMENTS

The development of these guidelines was led by David Foster, and they have been translated by Vicente Ponsoda. The Council of the ITC approved them in July 2014.

Why do we now need guidelines on the security of tests? The security of the measures has to do with the validity of the scores. If there are security problems, we do not know if the score we are assigning to the person being evaluated really reflects their level of knowledge, as it is assumed; it may reflect, in part, their ability to cheat on the answers, prior knowledge of some of the questions, that someone has helped them as they answered the test or afterwards, etc. Increasing the certainty that the scores on a test can be interpreted as the level of the person being evaluated in the construct of interest for the evaluation is a central issue, which provides the justification for inferences and the decisions based on them.

In the last two decades there has been a considerable increase in problems related to the security of measures worldwide. The two main reasons are, firstly, that computerised evaluation is increasingly widespread, often conducted by Internet and also often unattended (a type of assessment known in English as *Unproctored Internet Testing*), and, secondly, the proliferation of technologies that facilitate the recording, photographing and receiving of unauthorised information, etc., easily and with almost undetectable instruments. To these two reasons we should add a third. The more relevant the scores are for the person being evaluated, the more important the security problems are. The fact that evaluation through tests is increasingly common, particularly if they are tests whose results have important consequences for those being tested, is something that is possibly also related to the

major security problems observed in recent times in tests.

The guidelines are divided into three sections, related to: a) the development of the security plan, b) the inclusion of guidelines on the security of the evaluation process, and c) how to respond when a security breach occurs.

Before summarising the main guidelines of each section, it is worth making two comments. Firstly, the more important the consequences are for the person being evaluated, the teacher, school, etc., the more attention should be paid to the potential problems regarding security. When the consequences are barely important or not important at all, for example, when we ask our students to respond anonymously to a test we want to calibrate, no security problems are expected. Secondly, as noted above, computerised assessment has caused an increase in security issues, but paper and pencil tests may also present security problems. In fact, the guidelines are intended for both types of tests and may even be useful when the assessments are not standardised. They are also intended for international use; although they contain frequent warnings to consider the local regulations and usage when they are to be applied in each specific case.

DEVELOPING AND IMPLEMENTING A SECURITY PLAN

When responding to a test, there are many ways one can cheat in the responses. Some of the known ones are as follows: the person being evaluated obtains some or all of the test before answering it, someone helps the person being evaluated while he answers the test, the person being evaluated uses means that are not permitted during the examination (such as calculators, telephones, etc.), another person does the test instead of the person being evaluated, someone changes the answers of the person being assessed to increase their score, the person being evaluated copies the answers of another person being evaluated, etc. Yee and MacKown (2009) list thirty-seven ways to cheat in educational assessments. The list of known methods of test content theft is also considerable: theft of the books or digital documents containing the questions; recording the questions using, for example, a micro-camera; the employment of procedures that enable all of the information that reaches the computers in computerised tests to be recorded; the person being assessed memorises a few questions, others being evaluated memorise others, and with this coordinated effort they manage to steal all or most of the test content; the content can also be stolen by one of the many people who have access to the test during its preparation; etc. As noted, the aim of the guidelines is to reduce the risk of these problems arising in the evaluation program and, if they do occur, to provide guidelines on how to proceed.

The thirteen guidelines in this section indicate that the central element for controlling security is to develop a *security plan* that indicates a) who is responsible for each phase of test development, b) the rights and responsibilities of the person being assessed and how to put it down on record that the person being assessed knows them, c) what to do when there is a security breach, and d) the requirements to be met by the information and communication technology so that the conservation and transmission of the data is safe.

In a final guideline in this section, it is noted that the people involved in the development and application of the test must demonstrate that they know and accept the rules on the non-disclosure of their contents, who holds the copyright of the test, what behaviours are considered fraudulent and the possible consequences of such acts.

IMPLEMENTATION OF A SECURITY SYSTEM IN THE EVALUATION PROCESS USING TESTS

There are twenty-five guidelines in this second section and they indicate what to do to increase security at the different stages of development and implementation of the test. The main guidelines are briefly discussed below.

The people being evaluated must formally register and strict authentication processes must be applied, so that there is no doubt that the person doing the test is the one who registered to do it.

Some features of the application of the test are closely related to security and therefore it is recommended to prevent those being assessed from responding to the test more than once, that the test administration time is carefully studied to avoid giving extra time, and that the person being evaluated cannot go back to see the questions that she has already answered. We recommend using test and item formats that, whilst maintaining the psychometric quality, reduce exposure of the items or change their order of presentation, as with computerised adaptive tests. The number of times each item has been administered should be controlled; the bigger the item bank, the better; and the application of verification tests is also recommended when the computerised tests are applied via the Internet in unsupervised administrations.

The contents of the test must be protected during its development and distribution, and when it is being administered. As a security control during the

development and distribution process, the guidelines propose that only people who have to work with the items have access to them and for a limited time. To this end, strict access controls should be established and authorised individuals should sign nondisclosure agreements. Encryption is recommended as an additional security measure. As measures for increasing security during application, it is recommended to involve motivated vigilantes who preferably are not experts on the content being evaluated, as well as to install cameras to facilitate remote monitoring that can record any event of interest that occurs during the application of the test. If an anomaly is detected, it is recommended to respond quickly, temporarily or permanently interrupting the application of the test of the person being evaluated, confiscating the equipment used in the theft, if necessary, and preparing a report on the security incident. Another guideline suggests that the person being evaluated should know the security rules and the consequences of their violation prior to registering.

The test and item results should be evaluated regularly to check whether anyone has cheated and whether the items are known to those being evaluated before application of the test. Checks can be made for patterns of abnormal responses (where the difficult items are correct and the easy ones are wrong), patterns of abnormal response time (very short time), a high number of corrections on the answer sheet, high similarity among pairs or groups of people being evaluated (which may indicate copying), marked improvements when the test is repeated (this may indicate that the person has cheated), if there are changes in the distribution of item parameters (this may indicate that they have been leaked), if other types of items work as operational test items (for example, new items that have not previously been exposed), etc.

With regard to the obtaining and communication of the scores, it is recommended, especially for pencil and paper tests, to report that the score provided is "provisional" and that the "final" score will be communicated when the reports of irregularities that may have occurred have been considered.

It is useful to keep track of the Internet in order to be able to detect whether there has been disclosure of the test content. If so, the evaluation program must contact the person responsible for the website and request the removal of this content, announce the start of legal action, etc.

RESPONSE TO A FAILURE IN THE SECURITY SYSTEM

There are ten guidelines in this third part. They show what to do once there has been a failure in the security system.

As it was noted above, if the invigilators or supervisors of the test see that a person being assessed is cheating or stealing content, it is recommended that the application of the person being evaluated is suspended temporarily or permanently, and, if the law permits it, any instruments (camera, telephone, etc.) that have been used are seized, if necessary. Next, the failure in the security system must be thoroughly investigated to determine its extent and the magnitude of the damage, and, if necessary, the security plan must be revised.

Next we discuss some of the specific measures to be taken in relation to the test when a security breach has been detected. The test that has been leaked should be replaced as soon as possible. The scores that are known to be incorrect due to any type of fraud having been committed in the test must be cancelled and the person being evaluated must be regraded, either from his answers to the part of the test that was not leaked or by asking him to repeat the test or take an equivalent one.

Once the fundamentals of the security guidelines have been presented, it is important to bear in mind that the security of tests, as with security in other areas, is not a matter of all or nothing, but rather it is a continuum. Increased security involves a cost and every organisation has to find the point of balance. The purpose of these guidelines is not, therefore, to indicate what to do to avoid security problems, since there is no way of evaluating that completely eliminates the possibility that someone will cheat or steal the contents of the test. As in other areas where security must be ensured, it is likely that the security measures will always be a step behind the newly emerging anomalous behaviour (Foster, 2010). The purpose of these guidelines is to present the necessary process and best practice so that test assessments are more secure and preserve their value. The implementation of these guidelines will help prevent faults in the security system and minimise their consequences if they do occur.

In the last survey conducted in Spain by the COP to assess the attitude of Spanish psychologists to tests (Muñiz & Fernández-Hermida, 2010), the mean response to the question "The application of tests via the Internet opens up possibilities of fraud" was 3.78, on a scale from 1 (strongly disagree) to 5 (totally agree), above the neutral point 3, suggesting that licensed psychologists are aware

of the risks of the Internet administration of tests. In another survey (Ryan et al. 2015), applied to recruitment managers in American, European and Asian companies, respondents were asked about the security measures that they had applied in their assessments. In the case of unsupervised Internet assessments, only one measure was applied by over 50% of respondents: the strict control of the administration time (59% of respondents). The use of warnings about the evaluation including mechanisms for detecting anomalous behaviour was indicated by 40% and only 18% of respondents said they applied the verification tests, even though one of the guidelines expressly recommends this. In the case of monitored computerised assessments, the measures used by more than 50% of respondents were strict control of administration time (66%), use of passwords to access the contents of the test (58%), care in the surveillance tasks (56%) and not allowing copying of the contents of the test (55%). In the case of paper and pencil tests, the measures that over 50% of respondents adopt coincide with the four measures just presented and the percentages are very similar. The data from this survey, collected before the guidelines were approved, indicate that the security concern is real and steps have been taken in the past. The new guidelines should facilitate for professionals the adoption of the new measures and the development of an integrated system of security control that is more effective than the isolated measures that have been applied previously.

SOME FINAL THOUGHTS

We have presented the fundamental aspects of the ITC's statement on the use of tests in research and the guidelines on the quality control of measuring instruments, and everything related to the different factors to be considered in order to ensure the security of the assessment process. These guidelines are a great help in continuing to improve the use of tests because, as already noted in the introduction, is not enough that a test has adequate psychometric properties and the professionals who use it are well qualified, it also must be ensured that correct use is made of the test, and this is the aim of the guidelines described. The correct use of measuring instruments has important implications on two levels; from a professional point of view, it is essential that people are rigorously evaluated so that the decisions made by psychologists regarding different aspects of their life are appropriate and conform to the relevant deontological standards. From a scientific point of view, the use of appropriate measuring instruments is the only way to advance scientific psychology and to provide new tools for professional practice. A solid science of psychology with replicable results is only possible if the measuring instruments used have adequate metric properties. The replicability problems of current psychological research are due to various causes, but undoubtedly one of them is related to the measuring instruments used (Ioannidis et al, 2014; Koole & Lakens, 2012; Nosek & Lakens, 2014; Nosek et al., 2015). We hope that the guidelines presented help to continue to improve the use of tests both in research and applied contexts.

To facilitate an understanding of the framework of the guidelines presented in the current context of psychological evaluation, discussed below are the current perspectives and some of the pathways for the future development of the psychological evaluation, following the lines of those mentioned in previous works (Muñiz, 2012; Muñiz & Bartram, 2007; Muñiz, Elosua & Hambleton, 2013; Muñiz & Fernández-Hermida, 2010). The evaluation is constantly evolving, as is psychology itself, influenced by diverse factors, but undoubtedly the most powerful force driving the changes is the new information technology, particularly advances in computing, multimedia and the Internet. Some experts (Bennet, 1999, 2006; Breithaupt, Mills & Melican, 2006; Drasgow, Luecht & Bennet, 2006) believe that the new technologies are impacting on all aspects of the psychological evaluation, such as the design of the tests, the construction of the items, the presentation of the items, the scoring of the tests and remote assessment. All this is making the format and content of assessments change, the reasonable doubt emerging whether the paper and pencil tests, as we know them now, will be able to resist this new technological change. It is in this context of technological change that Psychology 2.0 emerges (Armayones et al., 2015), which aims to extend psychology through the facilities offered by the Internet and social networks. The evaluation cannot be indifferent to these new trends, new online psychometric approaches are emerging that are connected to the analysis of the large databases (big data) that are now available (Markovetz, Blaszkiewicz, Montag, Switala & Schlaepfer, 2014). For example, the potential benefits of using mobile phones as terminals for evaluation open up new possibilities for the future of psychometry (Armayones et al, 2015; Miller, 2012). Pioneering works, such as those by Kosinski, Stillwell and

Graepel (2013) successfully analyse the possibility of using Facebook "likes" as predictors of various human characteristics, including personality traits, which makes you wonder if our traces in social networks will someday soon replace the questionnaires and tests that we know today.

According to Professor Hambleton (2004, 2006), six major areas will attract the attention of researchers and practitioners in the coming years. The *first* is the international use of the tests, due to increasing globalisation and communication facilities, which raises a whole set of problems related to adapting tests from one country to another (Byrne et al, 2009; Hambleton et al, 2005; Muñiz et al, 2013). This internationalisation has highlighted the need for an overall evaluation framework to bring together good evaluation practices. Thus the International Institute for Standardization (ISO) has developed a new standard (ISO-10667), which contains the rules to be followed for the correct evaluation of people in work and organisational settings (ISO, 2011). The *second* is the use of new psychometric models and technologies to generate and analyse the tests. It is worth mentioning here all the new psychometry derived from Item Response Theory (IRT) models, which have managed to solve some problems that could not be solved within the traditional framework, but as always happens while solving problems, new problems arise that were not anticipated (Abad, Olea, Ponsoda & Garcia, 2011; De Ayala, 2009; Hambleton, Swaminathan, & Rogers, 1991; Muñiz, 1997; Van der Linden & Hambleton, 1997). The *third* is the appearance of new formats of items derived from the great advances in computer technology and multimedia (Irvine & Kyllonen, 2002; Shermis & Burstein, 2003; Sireci & Zenisky, 2006; Zenisky & Sireci, 2002). However, it is not about innovation for innovation's sake; before replacing the old formats with the new ones, it must be demonstrated empirically that they improve on the previous ones. The psychometric properties such as reliability and validity are not negotiable. The *fourth* area that will demand close attention is everything related to computerised tests and their relationship with the Internet. Deserving special mention in this area are the computerised adaptive tests that enable the test to be adjusted to the characteristics of the person being evaluated, without losing objectivity or comparability between people, which is opening very promising perspectives in evaluation (Olea, Abad & Barrada, 2010). Remote evaluation or tele-assessment is another line that is opening rapidly, raising, as we saw in the section on security guidelines, serious problems regarding the security of data and people, as it must be checked whether the person being evaluated is really who they claim to be, especially in the context of recruitment or tests with important implications for the future life of the person being evaluated (Bartram & Hambleton, 2006; Leeson, 2006; Mills et al., 2002; Parshall et al., 2002; Williamson et al., 2006; Wilson, 2005). Also deserving special mention on the subject of technology are developments concerning the automated correction of essays, which poses interesting challenges (Shermis & Burstein, 2003; Williamson, Xiaoming & Breyer, 2012). In *fifth* place, it is worth noting a field that may seem peripheral but is increasing in importance, which concerns the systems used to provide feedback on the results to the users and legitimately involved parties. It is essential that these individuals understand the results of the evaluations unequivocally, and the best way to ensure this is not obvious, especially if they have to be sent away for interpretation and explanation by the professional, as in many situations of recruitment, or educational evaluation (Goodman & Hambleton, 2004). Finally, it is likely that in the future there will be a great demand for *training* by the various practitioners involved in evaluation, not just psychologists, but also professionals such as teachers, doctors, nurses, etc. It is not about these practitioners being able to use and interpret psychological tests, but rather they will want information in order to understand and participate in the evaluation and certification processes being carried out in the workplace.

New forms of assessment are emerging, but psychometric tests will certainly continue to be fundamental tools, given their objectivity and economy in terms of means and time (Phelps, 2005, 2008). These are areas on which the evaluation activities of the not-too-distant future will most likely focus; it is not an exhaustive list by any means, but it gives some clues to guide people in the rapidly changing world of psychological assessment. The guidelines presented here have a transversal nature, being present in all of these areas outlined for the future, as in any of the circumstances mentioned there will always be tests used in research, it will always be necessary to carry out stringent quality control processes, and it will always be essential to ensure the security of the entire evaluation process.

REFERENCES

Abad, F.J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en ciencias sociales y de la salud [Measurement in social and health sciences]*. Madrid: Síntesis.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Armayones, M., Boixadós, M., Gómez, B., Guillamón, N., Hernández, E., Nieto, R., Pousada, M., & Sara, B. (2015). Psicología 2.0: Oportunidades y retos para el profesional de la psicología en el ámbito de la e-salud [Psychology 2.0: Opportunities and challenges for the psychology professional in the field of ehealth]. *Papeles del Psicólogo*, 36(2), 153-160.

Bartram, D. (1996). Test qualifications and test use in the UK: The competence approach. *European Journal of Psychological Assessment, 12*, 62-71.

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. & Coyne, I. (1998). Variations in national patterns of testing and test use: The ITC/EFPPA international survey. *European Journal of Psychological Assessment, 14*, 249-260.

Bartram, D. & Hambleton, R. K. (Eds.) (2006). *Computer-based testing and the internet: Issues and advances*. Chichester: Wiley.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and practice. 18*(3), 5-12.

Bennett, R. E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. In D. Bartram and R. K. Hambleton (Eds.), *Computer-based testing and the internet: Issues and advances*. Chichester: Wiley.

Breithaupt, K. J., Mills, C. N., & Melican, G. J. (2006). Facing the opportunities of the future. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the Internet* (pp. 219-251). Chichester: John Wiley and Sons.

Brennan, R. L. (Ed.) (2006). *Educational measurement*. Westport, CT: ACE/Praeger.

Byrne, B. M., Leong, F. T., Hambleton, R. K., Oakland, T., van de Vijver, F. J., & Cheung, F. M. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3*(2), 94-105.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Downing, S. M. & Haladyna, T. M. (Eds.) (2006). *Handbook of test development*. Mahwah, NJ: LEA.

Drasgow, F., Luecht, R. M. & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement*. Westport, CT: ACE/Praeger.

European Federation of Professional Psychologists' Associations (2005). *Meta-Code of ethics*. Brussels: Author (www.efpa.eu).

Fernández-Ballesteros, R., De Bruyn, E., Godoy, A., Hornke, L., Ter Laak, J., & Vizcarro, C. et al. (2001). Guidelines for the assessment process (GAP): A proposal for discussion. *European Journal of Psychological Assessment, 17*, 187-200.

Foster, D. F. (2010). Worldwide Testing and Test Security Issues: Ethical Challenges and Solutions. *Ethics & behavior, 20*(3–4), 207–228.

Goodman, D.P. & Hambleton, R.K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*, 145-220.

Hambleton, R. K. (2004). Theory, methods, and practices in testing for the 21st century. *Psicothema, 16*(4), 696-701.

Hambleton, R. K. (2006). *Testing practices in the 21st century*. Key Note Address, University of Oviedo, Spain, March 8th.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment*. London: LEA.

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

Hernández, A., Tomás, I., Ferreres, A. & Lloret, S. (2015) Tercera evaluación de tests editados en España [Third evaluation of tests published in Spain]. *Papeles del Psicólogo, 36*, 1-8.

Ioannidis, J. P. A., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18*, 235-241.

Irvine, S. & Kyllonen, P. (Eds.) (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

ISO (2011). *Procedures and methods to assess people in work and organizational settings (part 1 and 2)*. Geneva: ISO [Spanish version: Prestación de servicios de evaluación: procedimientos y métodos para la evaluación de personas en entornas laborales y organizacionales (partes 1 y 2). Madrid: AENOR, 2013].

Joint Committee on Testing Practices. (2002). *Ethical principles of psychologists and code of conduct*. Washington DC: Joint Committee on Testing Practices.

Koocher, G. & Keith-Spiegel, P. (2007). *Ethics in psychology*. New York: Oxford University Press.

Koole, S. L. & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science, 7*, 608-614.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behaviour. *Proceedings of the National Academy of Sciences (PNAS)*, 110(15), 5802-5805.

Leach, M. & Oakland, T. (2007). Ethics standards impacting test development and use: A review of 31 ethics codes impacting practices in 35 countries. *International Journal of Testing, 7*, 71-88.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing, 6*, 1-24.

Lindsay, G., Koene, C., Ovreeide. H., & Lang, F. (2008). *Ethics for European psychologists*. Gottingen and Cambridge, MA: Hogrefe.

Markovetz, A., Blaszkiewicz, K., Montag, C., Switala, C, & Schlaepfer, T. E. (2014). Psycho-Informatics: Big Data shaping modern psychometrics. *Medical Hypotheses*, 82(4), 405-411.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, 7(3), 221-237.

Mills, C.N., Potenza, M.T., Fremer, J.J., & Ward, W.C. (Eds.) (2002). *Computer-based testing: Building the foundation for future assessments*. Hillsdale, NJ: LEA.

Moreland, K. L., Eyde, L. D., Robertson, G. J., Primoff, E. S., & Most, R. B. (1995). Assessment of test user qualifications. *American Psychologist, 5*(1), 14-23.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems [Introduction to Item Response Theory]*. Madrid: Pirámide.

Muñiz, J. (2012). Perspectivas actuales y retos futuros de la evaluación psicológica [Current perspectives and future challenges of psychological assessment]. In C. Zúñiga (ed.), *Psicología, sociedad y equidad [Psychology, society and equality]*. Santiago de Chile: Universidad de Chile.

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

Muñiz, J., Elosua, P., & Hambleton, R.K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición [Guidelines for the translation and adaptation of tests: Second Edition]. *Psicothema, 25*(2), 151-157.

Muñiz, J. & Fernández-Hermida, J. R. (2010). La opinión de los psicólogos españoles sobre el uso de los tests [The opinion of Spanish psychologists on the use of tests]. *Papeles del Psicólogo, 31*, 108-121.

Muñiz, J., Prieto, G., Almeida, L., & Bartram, D. (1999). Test use in Spain, Portugal and Latin American countries. *European Journal of Psychological Assessment, 15*(2), 151-157.

Muñiz, J., Fernández-Hermida, J.R., Fonseca-Pedrero, E., Campillo-Álvarez, A. & Peña-Suárez, E. (2011). Evaluación de tests editados en España [Review of tests published in Spain]. *Papeles del Psicólogo, 32*, 113-128.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T. A., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Levy Paluck, E., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., & Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*, 1422-1425.

Nosek, B. A. & Lakens, D. (2014). Registered reports. A method to increase the credibility of published results. *Social Psychology, 45*, 137-141.

Olea, J., Abad, F., & Barrada, J. R. (2010). Tests informatizados y otros nuevos tipos de tests [Computerised tests and other new types of test]. *Papeles del Psicólogo*, 31(1), 94-107.

Papeles del Psicólogo (2009). *Número monográfico sobre Ética Profesional y Deontología [Special issue on professional ethics and deontology]*. Vol. 30, 182-254.

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Phelps, R. (Ed.) (2005). *Defending standardized testing*. London: LEA.

Phelps, R. (Ed.) (2008). *Correcting fallacies about educational and psychological testing*. Washington: APA.

Ponsoda, V. & Hontangas, P. (2013). Segunda evaluación de tests editados en España [Second evaluation of tests published in Spain]. *Papeles del Psicólogo, 24*, 82-90

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

Ryan A.M., Inceoglu, I., Bartram, D., Golubovich, J., Grand, J., Reeder, M., Derous, E., Nikolaou, I. & Yao, X. (2015). Trends in testing: Highlights of a global survey. In Nikolaou, I. & J. Oostrom (Eds.). *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice.* Psychology Press-Taylor & Francis.

Shermis, M. D. & Burstein, J. C. (Eds.) (2003). *Automated essay scoring*. London: LEA.

Simner, M. L. (1996). Recommendations by the Canadian Psychological Association for improving the North American safeguards that help protect the public against test misuse. *European Journal of Psychological Assessment, 12*, 72-82.

Sireci, S., & Zenisky, A. L. (2006). Innovative items format in computer-based testing: In pursuit of construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Hillsdale, NJ: LEA.

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

Williamson, D.M., Mislevy, R.J. & Bejar, I. (2006). *Automated scoring of complex tasks in computer based testing*. Mahwah, NJ: LEA.

Williamson, D.M., Xi, X., & Breyer, J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13.

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

Yee, K., & MacKown, P. (2009). Detecting and preventing cheating during exams. In T. Twomey, H. White, & K. Sagendorf (Eds.), Pedagogy not policing: Positive approaches to academic integrity at the University (pp. 141 - 148). Syracuse: The Graduate School.

Zenisky, A.L. & Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337-362.