Papeles del Psicólogo, 2010. Vol. 31(1), pp. 85-96
http://www.cop.es/papeles

PERFORMANCE ASSESSMENT

Rosario Martínez Arias
Complutense University of Madrid

Assessment practices have gradually shifted from almost exclusively objectively-scored multiple-choice test items to the use of a mixture of formats, including performance assessments. The purpose of this article is to provide an overview of the concept, design, use and psychometric characteristics of performance assessment. The article is divided into five sections. It begins with the concept and rationale for the use of performance assessment. Section 2 presents the main uses of these tests. Section 3 covers some questions related to their design and scoring. Section 4 deals with some issues related to psychometric characteristics. Finally, Section 5 evaluates performance assessment tests, indicating their main strengths and weaknesses and future research needs. Continued work is needed on measurement models and methods that can improve the generalizability and evidence of validity of performance assessments. Computer applications can make a contribution to practical issues.
Key words: *Performance assessment, Assessment centers, Scoring rubrics, Generalizability, Evidence of validity.*

Las prácticas de evaluación han evolucionado desde el uso casi exclusive de tests formados por ítems de elección múltiple a la combinación de formatos múltiples, incluyendo tareas de desempeño. El objetivo del artículo es proporcionar una visión del concepto, diseño, uso y características psicométricas de los tests de desempeño. Comienza con el concepto y la justificación de su uso. En la sección 2 se presentan los principales usos actuales de este tipo de tests. La sección 3 describe algunos aspectos relativos al diseño y puntuación. La sección 4 muestra algunas cuestiones relativas a las características psicométricas. La sección 5 concluye con una valoración de los tests de desempeño, presentando sus principales fuerzas y debilidades, así como las necesidades de futuras investigaciones. Se necesita un esfuerzo continuado en modelos y métodos de medida que permitan mejorar la generalizabilidad y las evidencias de validez de los tests de desempeño.
Palabras clave: *Test de desempeño, Centros de evaluación, Guías de puntuación, Generalizabilidad, Evidencias de validez.*

THE CONCEPTOF PERFORMANCE ASSESSMENT

The standardized test is widely considered, even among psychology professionals, to be synonymous with the multiple-choice or single constructed response test. Such a notion is understandable in view of the fact that these formats have dominated the field of tests measuring intelligence, aptitudes and academic performance for many years – and for good reasons, related above all to the range of content they cover and the ease with which they can be marked and scored. However, the standardized test label can also be applied to other formats that fulfill all the requirements of a test and which can show adequate psychometric properties. Among these would be those we refer to here under the heading of *performance assessment*[1], increasingly employed in psychological and educational assessment.

Table 1 shows a classification of different types of standardized test format, which can be situated along a series of continuums (Gronlund, 2006).

Among the formats considered, those generally regarded as performance assessment are essays, projects, simulations and work samples. As it can be seen, these formats are closer to the extremes characterized by greater authenticity and cognitive complexity, more in-depth coverage and response structured by the respondent him/herself. They also tend to be more expensive.

Given the wide range of formats performance assessments can take, a possible definition that takes into account such diversity would be as follows: "performance assessments are standardized assessment procedures in which respondents are required to carry out tasks or processes in which they demonstrate their ability to apply knowledge and skills to actions in simulated or real-life situations".

Correspondence: Rosario Martínez Arias. *Departamento de Metodología de las Ciencias del Comportamiento. Universidad Complutense de Madrid. E-mail: rmnez.arias@psi.ucm.es*

.....

[1] The term "performance assessment" originated in the fields of educational assessment and professional certification; however, this type of test has been used in psychology for many years, especially in the field of personnel selection. Although the actual expression "performance assessment" is not used, the simulation tasks and work samples employed in *assessment centers* display all the characteristics of such tests, insofar as they require responses which place emphasis on the examinee's performance and require systematic methods for their scoring. With the advent of new technologies their use has been extended to other areas of the discipline, such as clinical psychology and neuropsychology

Such assessments can involve activities as diverse as writing an essay, playing a musical composition, giving an oral presentation, diagnosing a standardized patient, planning the day's activities or proposing a solution to a business problem. In all cases the respondent has to produce something during a given period of time and the processes or products are assessed in relation to established performance criteria.

The definition is a comprehensive one in that it includes the two broad groups into which such definitions are usually divided: those which place the emphasis on the response format and those which give more importance to the similarity between the required response and the criterion of interest (Palm, 2008). In this latter group, the majority focus on the examinee's performance (Stiggins, 1987).

Some, indeed, go beyond the response format, insisting on the *authenticity* and *simulation* characteristics of the criterion situation. Thus, Fitzpatrick and Morrison (1971) define them as "those tests in which a criterion situation is simulated more faithfully and comprehensively than in the usual paper-and-pencil tests" (p.268). According to Kane, Crooks and Cohen(1999), they "represent a sample of the subject's performance in some domain, the resulting scores being interpreted in terms of typical or expected performance in that domain…, their defining characteristic being the high level of similarity between the type of performance observed and that of interest" (p.7). In a similar line is the definition found in the *Standards for Educational and Psychological Tests* (American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME), 1999), according to which, "performance assessments *emulate* the context or conditions of application of the knowledge and skills they seek to assess"(p.137).

This insistence on the emulation of the performance of interest leads to some confusion with so-called *authentic assessment* (Wiggins, 1989), which shares many characteristics with performance assessment and constitutes one of its forms, but involves other aspects that go beyond those demanded by these tests.

Also frequently highlighted is the cognitive complexity involved, given that respondents are required to use higher-order strategies, such as planning, task structuring, obtaining information, constructing responses and explaining the process, combining knowledge and information(Ryan, 2006).

Performance assessments are generally classified according to what they assess, and in this sense we tend to speak of *products*, or the results of the task, and *performances*, which are the processes followed by the examinee in order to reach the solution. Typical examples of the first aspect are written essays, lab reports, artistic performances, and so on. Among the second aspects would be oral presentations and demonstrations. In the majority of cases there is a combination of processes and products.

USE OF PERFORMANCE ASSESSMENTS

Performance assessments are far from being a new concept; Madaus and O'Dwyer (1999) situate their origins as early as 210 BC, during the Han Dynasty in China. Similar types of assessment were used by the guilds during the Middle Ages and in the universities for assessing students. Within work psychology there is a long tradition of their use in the army, and for more than 60 years they have been employed in so-called *Assessment Centers*, known today as *Assessment and Development Centers* (Thornton, & Rupp, 2006), which use samples of work and simulation exercises to assess individuals in competencies that are difficult to measure using conventional tests. Their use in Britain by the *War Office Selection Boards* for the selection of senior officers dates back to 1942, and this practice soon spread to the United States and elsewhere, especially the German-

TABLE 1
TEST FORMAT CONTINUUMS

Work samples	Simulations	Projects	Essays	Short answer	Multiple-choice	True/False
Most authentic	←				→	Least authentic
Cognitively most complex	←				→	Cognitively least complex
In-depth coverage	←				→	Coverage of content
Response structured by examinee	←				→	Response structured by test
Highest cost	←				→	Lowest cost

speaking countries. They have also been used in relation to management positions since the 1950s, though today their use extends to many more types of post (Thornton, & Rupp, 2006).

In educational assessment, strong criticism of the multiple-choice format in the 1960s and 70s led to the inclusion of tasks based on performance assessment. During the 1990s the exclusive use of multiple-choice formats gave way to mixed formats for the assessment of performance, featuring written essays, problem-solving sequences, oral presentations, and even student *portfolios* (Hambleton, 2000). The reasons for the change are diverse, but basically have to do with the limitations of multiple-choice tests for achieving certain educational objectives: 1) the assessment of high-level cognitive abilities; 2) the evaluation of life-long learning skills (independent thinking, flexibility, etc.); 3) the assessment of strategies for solving problems and dealing with difficulties; 4) the alignment of skills and abilities with competencies that are important for life and with realistic contexts, and 5) the integration of assessment and instruction in line with theories of learning and cognitive psychology. Such objectives are included in educational reforms that place the emphasis on the teaching of higher-order cognitive skills (Linn, 1993a) and the bond between assessment and instruction, assessment being considered a valuable instrument for the improvement of instruction and learning (Frederiksen, & Collins, 1989; Stiggins, 1987). They form part of efforts to address the widely discussed issue of the "dumbing-down" of curricula resulting from the use of multiple-choice tests, in the belief that assessment determines what teachers teach and students learn (Wiggins, 1989).

The progress made by cognitive psychology was an important factor contributing to the inclusion of performance-based tasks in assessment processes. In 1998 the National Research Council (NRC) *Board on Testing and Assessment* formed a committee of 18 experts chaired by Pellegrino and Glaser with the aim of bridging the gap between advances in cognitive psychology and methods of educational measurement. The end product was the excellent work "*Knowing What Students Know: The Science and Design of Educational Assessment*" (NRC, 2001), which highlights the limitations of traditional tests for measuring the knowledge and complex skills required by new performance criteria and the scarce validity of the inferences derived from their scores. The committee developed a theoretical framework for assessment, the *Assessment Triangle*, based on the idea that assessment is a *process of evidence-based reasoning* (Mislevy, 2006; Mislevy, Steinberg, & Almond, 2002; Mislevy, Wilson, Ercikan, & Chudowsky, 2003) with three supporting pillars: a) a representative model of knowledge and the development of competencies, b) tasks or situations that permit the observation of students' performance, and c) interpretation methods for making inferences.

Furthermore, the advent of computers opened up the possibility of using new item and response formats, facilitating both the administration and the scoring of such tasks (Drasgow, Luecht, & Bennett, 2006; Zenisky, & Sireci, 2002).

Today, performance assessments are present in the majority of large-scale assessments, generally accompanied by items with structured format. In the United States they started to be included in the *National Assessment of Educational Progress* (NAEP) during the 1990s, and today many states use performance assessments in their annual programs of tests. They are also included in all large-scale international assessments, such as *Trends in International Mathematics and Science Study*, TIMSS (Arora, Foy, Martin, & Mullis, 2009), and in the PISA program (OECD, 2007). In Spain they have been incorporated into the diagnostic tests developed by the *Instituto de Evaluación*, or Assessment Institute.

In the field of work, performance assessments are strongly represented in professional accreditation, especially for the practicing of medicine and law. An example of the former would be the *United States Medical Licensure Examination* (USMLE, 2009), and of the latter, the *Multistate Performance Test*, employed in 30 US states (National Conference of Bar Examiners & American Bar Association, 2005)[2].

A large part of the tasks in these tests are similar to those used in assessment centers for personnel selection. In these systems the type of task adopts multiple forms, though the most common are the following: in-tray tests, role-play in interactions, analysis of written cases from the organization, oral presentations, leadership in group discussions, search for relevant facts on the basis of oral

[2] Detailed descriptions of the performance assessments used in diverse types of professional accreditation can be found in Johnson, Penny and Gordon (2009).

presentations, business games, and combinations of several tasks or exercises. A description of assessment center tasks can be found in Thornton and Rupp (2006).

Examinees' behavior is assessed according to dimensions relevant to the jobs in question, the number and type of which differ depending on the assessment center's objective (Thornton, & Rupp, 2006). Some are common to the majority of centers and similar to those used in certifications: problem-solving, oral communication, leadership, conflict management, search for information, planning and organization, cultural adaptability, generation of solutions, use of resources, and so on (Arthur, Day, McNelly, & Edens, 2003; Brummel, Ruth, & Spain, 2009).

THE DEVELOPMENT, ADMINISTRATIONAND SCORING OF PERFORMANCE ASSESSMENTS

Performance assessments must ensure that the exercises or tasks are standardized, valid, reliable, equitable and legally defensible. To achieve this, their development process should be in line with standards and guidelines for the construction and use of tests such as the *Standards for educational and psychological tests* (AERA et al., 1999). In the case of assessment center exercises, they should also be in accordance with some specific guidelines, such as the *Principles for the validation and use of personnel selection procedures* (Society for Industrial and Organizational Psychology, 2003) and the *Guidelines and Ethical Considerations for Assessment Center Operations* (International Task Force on Assessment Center Guidelines, 2000).

The development process begins with the *definition* of the framework, which involves the description of the construct or tasks, the purpose of the assessment and the inferences to be made from the scores. The conceptual framework guides the development of the *specifications*, which reflect the content, the processes, the psychometric characteristics of the tests and other information pertinent to the assessment. Two approaches can be followed, focused on either the construct or the task, though the former is recommended (Messick, 1994). The construct guides the appropriate representation of the domain, the selection of the tasks, the scoring criteria and the detection of possible irrelevant variance. Patz (2006) provides a good description of the development of an assessment in sciences. In assessment centers, the framework of definition of the constructs or competencies is derived from a rigorous analysis of the job in question (Thornton, & Rupp, 2006).

For appropriate standardization it is necessary to determine the conditions of administration that permit comparability of the scores (AERA et al., 1999). *Scoring rubrics* are drawn up which set the time frames, items or tasks, equipment and materials, and application instructions (Cohen, & Wollack, 2006).

The key to the success of these tests, and one of the most controversial aspects, is the correct *assignment of scores* to the tasks carried out. For this purpose, *scoring rubrics* are drawn up, which establish the criteria for rating responses and a procedure for scoring them (Clauser, 2000). They must be clear and comprehensive, and illustrated with examples of typical responses (Welch, 2006). Their objective is to ensure that scores are consistent and invariant across raters, tasks, locations, occasions and other conditions. Combined with the appropriate training of raters, they make it possible to attain adequate levels of reliability.

There are two types of rubric, *holistic* or *global*, and *analytical*. In the global type, raters make a single judgment on the quality of the process or product, assigning a score based on *anchored* descriptions for the different levels. In the analytical type, the performance descriptions are separated into parts (aspects, assessment criteria, dimensions, domains, etc.). In addition to the categories of the rubrics, exemplar responses are included for operationalizing each of the assessment criteria, called *anchors*, or points of reference.

An analytical rubric specifies detailed features or aspects of the responses and the number of points that should be awarded to each one, allowing weighting. The different features are usually scored by means of Likert-type scales with several levels. Assessment centers use a similar procedure to that applied in analytical rubrics, known as *Behaviorally Anchored Rating Scales* (BARS), which includes exemplar ("anchored") descriptions of behaviors and permits the rating of each dimension on scales which generally have five points.

A variation of the analytical scoring system is that of *checklists* for behaviors, on which each aspect is rated Yes or No, according to whether it is present or absent. This is the customary procedure in medical and legal accreditation, and is sometimes used in assessment centers instead of BARS.

When the tasks are based on cognitive theories of learning within a domain, the scores may reflect criteria of progression in learning (Wilson, 2005).

The choice of one form or another largely depends on the construct, the purpose of the assessment, whether it is a process or a product that is being assessed, and the inferences that will be drawn from the scores. The number of categories or points on the scales depends on the facility of differentiation and discrimination. According to Lane and Stone (2006), they should have sufficient categories for differentiating between levels of performance, but not so many as to make the differentiation difficult.

Among the most widely explored aspects are the relative merits of the two scoring systems, analyzed on the basis of inter-judge reliability. So far, no procedure has emerged as clearly superior in all situations. It would appear that holistic rubrics are more affected by sources of bias in raters than the analytical type, and that checklists of behaviors improve inter-rater agreement. Johnson et al. (2009) and Arter and McTighe (2001) recommend holistic rubrics for relatively simple tasks, such as those included in large-scale assessments. Analytical rubrics are more appropriate for complex tasks with multiple elements, as are the cases of licenses and certifications and of assessment centers (Welch, 2006).

The considerable costs of scoring these tests have led to the development of some computerized systems for this purpose (Bennett, 2004; Livingston, 2009; Williamson, Mislevy, & Bejar, 2006). Their implementation involves identifying a large number of typical responses rated by experts, which represent the total range of scores on the scale, and then using algorithms for obtaining the scores that emulate human raters (Williamson et al., 2006).

Another crucial aspect is the training of the raters with whom it is attempted to reach adequate levels of agreement, correcting their biases. The most common biases are summarized in Table 2, adapted from Johnson et al. (2009).

A widely used training procedure involves the inclusion of protocols previously corrected by experts, which permit the detection of raters with bias and monitoring by experienced raters.

PSYCHOMETRIC CHARACTERISTICS OF PERFORMANCE TESTS

Performance assessments must meet psychometric criteria in the same way as any other assessment procedure (Kane, 2004), and in this regard, different models of test theory are used. Some specific characteristics demand the use of models more advanced than Classical Test Theory (CTT), such as Generalizability Theory (GT) and Item Response Theory (IRT). New conceptions of validity also lead to some differences with respect to traditional approaches (see the articles by Muñiz (2010) on test theory and by Prieto and Delgado (2010) on reliability and validity, both in this same issue).

Below we briefly review some psychometric aspects of performance assessments: the treatment of measurement errors and score consistency (reliability), procedures for obtaining ability estimations and evidence of validity. This conventional classification is difficult to apply to these tests, since the generalizability of the scores is often treated as one of the aspects of validity (Brennan, 2000a; Kane, Crooks, & Cohen, 1999; Miller, & Linn, 2000; Messick, 1996).

TABLE 2
MOST COMMON TYPES OF RATER BIAS

Types of bias	Rater's tendency to…
Appearance	Score based on the looks of the response
Central Tendency	Assign scores primarily around the scale midpoint
Clashing standards	Score lower because his or her personal grading standards conflict with standards expressed in the rubric
Fatigue	Allow scores to be affected by being tired
Halo effect	Score higher because some positive aspect of a performance positively influences a rater's judgment
Handwriting	Allow handwriting to influence his or her scores
Item-to-item carryover	Score higher a response because an examinee's performance on the preceding item was exemplary
Language	Score according to language usage when other dimensions are the focus of the rating
Length	Score lengthy responses higher
Leniency/severity	Score too easily or harshly
Repetition factor	Lower a score because he or she has read about a topic or viewed a response repeatedly
Sudden death	Score lower due to some aspect of the performance that provokes a negative rater response
Test-to-test carryover	Score lower a response that meets the pre-stated expectations, but the response appears somewhat lacklustre as compared to exemplary responses that preceded it
Trait	Focus on one aspect (i.e., trait), such as conventions, and give too much weight to this trait in arriving at a score

Reliability and consistency of scores

Reliability can sometimes be dealt with using CTT (Johnson et al., 2009), but it is usually necessary to employ GT. This latter model, systematized by Cronbach, Gleser, Nanda and Rajaratnam (1972), made little impact in the field of test construction until the advent of performance assessments, in which its use has become generalized. GT is an extension of CTT that uses Analysis of Variance models (components of variance) which permit the simultaneous estimation of the effects of different sources of variability or error (*facets*) on scores. The facets most commonly considered are tasks and raters, though some studies include occasions, administration and test format. The model allows the analysis of the principal effects of each facet, as well as their interactions with the respondent and with one another. There are two types of GT study, G (Generalizability) and D (Decision). In the former, the aim is to estimate the relative contribution of each facet and of its interactions in relation to the error variance, and these estimations permit the optimization of the measurement process, determining the optimum number of tasks, raters, etc. in each application for reducing error. In D studies the objective is to calculate the *generalizability coefficient* sunder the specific measurement conditions used; these studies can be of two types, according to whether the decisions are *absolute* or *relative*. The possibility of breaking down the error variance into different sources is what makes GT essential in performance assessments. The reliability improves when studies are carried out to determine the necessary numbers of tasks and raters.

Reasons of space prevent us from going into more detail in our description of GT. The issue is dealt with comprehensively in Brennan (2000b), while Martínez Arias, Hernández Lloreda and Hernández Lloreda (2006) offer a useful summary and Prieto and Delgado (2010), in this special issue, describe its principal characteristics.

The sources of error most widely studied are those related to the *task* and the *rater*. It has been found that the most critical effects are those of the tasks, given the small number of them that can be included for each ability or competency, with low consistency between tasks and interaction effects with examinees (Lane, & Stone, 2006).

The effect of the raters is important, both as a principal effect and in interaction with tasks and respondents. In written assessments, moderate and high correlations among judges have been found, ranging from 0.33 to0.91 (Lane, & Stone, 2006), and they are even better in the case of medical accreditations, with values of between 0.50 and 0.93 (van der Vleuten, & Swanson, 1990). As regards the type of competency assessed, there is greater consistency for assessments of sciences and mathematics compared to those of written work (Shavelson, Baxter, & Gao, 1993).

In general, it can be said that task variability contributes more to error than the rater in the majority of fields (Lane, & Stone. 2006; Shavelson et al., 1993).

Some IRT models developed within the framework of the Rasch model (Adams, Wilson, & Wang, 1997) permit the incorporation of the effects of the rater on the scores.

Although tasks and raters are the most widely studied sources of variability, the effects of other facets have also been explored: occasions (measurement time points), assessment format and raters' committee. An important facet is occasion (Cronbach, Linn, Brennan, & Haertel, 1997; Fitzpatrick, Ercikan, Yen, & Ferrara, 1998), especially in periodical assessments in which changes are examined and different raters award scores.

Estimation of competency or ability

To obtain estimations of examinees' competency or ability it is customary to use as a framework the IRT models for ordered polytomous responses (Abad, Ponsoda, & Revuelta, 2006). Recent advances in the field of multidimensional IRT models (*Multidimensional Item Response Theory*, MIRT) have made it possible to work with the complexity of these assessments, in which it is difficult to obtain the assumed unidimensionality (Gibbons et al., 2007; Reckase,2009).

A common problem is the combination of different response formats in the same test. The use of IRT models with specialized software for polytomous models makes it possible to obtain single estimators of skills or traits in these conditions.

In relation to the estimation of scores there arises the problem of *equating*, when different sets of items are used, in the same assessment or at different time points to assess changes. The characteristics of these tests lead to particular problems for the application of equating techniques in the strict sense (Kolen, & Brennan, 2004), so that weaker forms, such as calibration, prediction or moderation, often have to be used (Linn, 1993b). The main problems are the frequent multidimensionality, the difficulty of finding common anchoring items, the polytomous nature of the items and dependence among items (Muraki, Hombo, & Lee, 2000), as well as effects of

the rater (Kolen, & Brennan, 2004). Multigroup IRT models are often employed to deal with the situation (Bock, Muraki, & Pfeiffenberger, 1988). Reckase (2009) proposes some procedures in the context of multidimensional models. A recent treatment of these problems can be found in Dorans, Pommerich and Holland (2007).

Evidence of validity of performance assessments

The definition of validity of performance assessment scores is that provided in the *Standards for Educational and Psychological Tests* (AERA et al., 1999), and similar to that of other types of standardized test, with construct validity as a unifying concept. The article by Prieto and Delgado (2010) in this special issue deals with the definition and types of evidence. In performance assessments other aspects are often mentioned, such as authenticity, meaningfulness for examinees (Linn, Baker, & Dunbar, 1991) and systemic validity (Fredericksen, & Collins, 1989). Messick (1996) considers these aspects within the contexts of construct representation (authenticity) and of substantive and consequential aspects of validity (meaningfulness and systemic validity).

Below we briefly review evidence of validity, with some considerations about bias and equity, which can also be dealt with in the contexts of irrelevant variance for the construct and of consequences.

Evidence of content validity

Performance assessments are more prone than conventional tests to the two major threats to content validity identified by Messick (1989, 1996): *under-representation of the construct* and *irrelevant variance*. The former is usually due to the small number of items they include; the latter has multiple sources: choice of topic by examinees, the tendency of raters to focus on irrelevant aspects or biases (Messick, 1994, 1996; see Table2), automated marking procedures (Lane, & Stone, 2006) and examinees' motivation, especially in low-stakes assessments (DeMars, 2000; O'Neil, Subgure, & Baker, 1996).

Evidence of validity from response processes

Messick (1996) highlights "the need to obtain empirical evidence of the processes set in motion by examinees when they perform the task" (p.9). Given the expectations resting on these tests that they will assess higher-order cognitive processes, it is important to ascertain that they actually do so (Hambleton, 1996; Linn et al., 1991).Up to now research has been scarce, and the results somewhat inconsistent (Ayala, Shavelson, Shue, & Schultz, 2002). Some developments inspired in the *Latent Trait Logistic Model* (Fischer, 1973), such as those of Embretson (1998) and Gorin and Embretson (2006), are promising in this regard. Adams, Wilson and Wang (1997) developed a multidimensional version, appropriate for these types of test.

An interesting theoretical framework is that of the *Assessment Triangle*, mentioned in the second section of this article. In assessments of learning results in the educational context, the *developmental assessment* approach can also provide support for this type of validity (Briggs, Alonzo, Schwab, & Wilson, 2006; Wilson, 2005).

Structural

According to AERA et al. (1999), "the analysis of a test's internal structure can indicate the extent to which the relations between its items and components are appropriate for the construct on which the interpretations of the scores are based" (p.13).

The assessment of dimensionality usually takes place by means of factor analysis techniques. There are few published works on the factor structure of performance assessments in education, the reasons for which are diverse: 1) the complexity of the stimuli leads to the recommendation of content analysis and substantive analysis (Ackerman, Gierl, & Walker, 2003); 2) the scoring schemes can affect the dimensionality, sometimes leading to multidimensionality, and 3) different points on the rating scale can reflect different combinations of skills (Reckase, 1997). Advances in the field of multidimensional IRT models (Gibbons et al., 2007; Reckase, 2009) may help to provide structural evidence. In the context of assessment centers more work has been done on this aspect, with contradictory results. Rupp et al. (2006) found evidence of clear dimensions, but other authors have questioned them (Lance, 2008).

External

In the words of AERA et al. (1999), "analysis of the relations between test scores and external variables is another important source of evidence of validity" (p.13).Such evidence is obtained by examining the patterns of empirical correlations according to the theoretical expectations or hypotheses of the construct.

Messick (1996) stresses the importance of evidence of *convergent* and *discriminant* validity from multitrait-multimethod (MTMM) matrices. "*Convergent evidence* signifies that the measure in question is coherently related to other measures of the same construct as well as to other variables that it should relate to on theoretical grounds. *Discriminant evidence* signifies that the measure is not unduly related to exemplars of other distinct constructs" (Messick, 1996, p.12).

It should be shown that the variance due to the construct is considerably greater than the variance of the method or the tasks. In the educational context there is little published research on such evidence, but it has been widely studied in assessment centers, where low evidence has tended to be found, since the proportion of variance related to the construct is usually greater than that related to the construct (Lance, 2008). Nevertheless, Rupp, Thornton and Gibbons (2008) attribute these results to methodological deficiencies in the design of the multitrait-multimethod matrices.

As regards the evidence related to external criteria, the majority of research has been carried out in the assessment center context. In a meta-analysis, Arthur et al. (2003) found correlations of between0.25 and0.39, depending on the types of competencies. Salgado and Moscoso (2008), reviewing the reliability and effective validity (with correction of bias due to lack of criterion reliability and range restriction) of various selection instruments, found reliability and validity coefficients of 0.70 and 0.37, respectively, for assessment center simulations, the latter being lower than those found for other procedures (tests of general aptitude and reasoning, tests of job knowledge and structured behavioural interview). These data raise doubts about the utility of these procedures with respect to others which, moreover, are more economical.

Consequences of test use

This aspect of construct validity has to do with the desirable and undesirable consequences of test use and their impact on the interpretation of scores (Messick, 1996). Such evidence has been studied in the educational context, in which it constitutes one of the arguments most often put forward for the use of performance assessments. Among the positive consequences for examinees are motivation, learning and the application of what they have learned.

Up to now, research in this area has been scarce. Stecheret al. (2000), in a survey of teachers, found that two-thirds of 4th to 7th-grade teachers said state standards and performance tests influenced their teaching strategies. The impact of the Maryland State Performance Assessment Program was examined by Lane and cols. (Lane, Parke, & Stone, 2002; Parke, Lane, & Stone, 2006), who found that both management personnel and teachers considered the assessment to have brought about positive changes in teaching and in assessment practices in the classroom. However, this result may derive from the consequences of the assessment for the schools in question (accountability).

Bias and equity

Bias is generally understood as "the differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers" (Cole, & Moss, 1989, p.205). To avoid bias, it is recommended to use techniques for detecting *differential item functioning*, which permit the identification of tasks or items that may contribute to bias. For a more detailed description, see the article by Gómez Benito, Hidalgo and Guilera (2010) in this issue.

Few studies have been carried out on differential item functioning in performance assessments (Lane, & Stone, 2006). The majority of research has been confined to analysis of the differences between groups. With regard to written essays, differences are found between males and females, in favour of the latter (Ryan, & DeMark, 2002), and between ethnic groups (Engelhard, Gordon, Walker, & Gabrielson, 1994). In studies more appropriate for the analysis of bias, non-parallel differences have been found between the performance assessment and multiple-choice formats (Livingston, & Rupp, 2004). When males and females show similar results in multiple-choice formats, females are better in those of constructed response; when the two sexes are similar in constructed response formats, males are better in those of multiple-choice.

It is considered that performance assessments can involve more irrelevant factors, which can lead to differential functioning (Penfield, & Lamm, 2000). Its detection is more difficult in these types of assessment, given the above-mentioned problems in relation to equating.

CONCLUSIONS

Today, performance assessments form part of the

repertoire of assessment techniques, and are increasingly widely used. They have generated considerable expectations as a result of their apparent validity and their potential advantages: greater authenticity through the emulation of real situations, the possibility of measuring abilities and skills that are difficult to assess with other formats, the measurement of processes as well as products, their great value in educational and training and their potential for detecting progress in learning. All of this makes them essential in the field of assessment, normally in combination with tests or tasks in more traditional formats. Moreover, the innovations derived from the use of new technologies assist their application, paving the way to the assessment of new competencies and dimensions.

Nevertheless, and in spite of their undeniable advantages and widespread use, they still present numerous challenges for psychometric research. Their chief limitations are as follows:

- ✔ Difficulties in adequately representing the domain due to the limited number of tasks that can be included.
- ✔ Problems of generalizability, resulting above all from variance due to the tasks and to the interaction of the tasks with examinees and raters.
- ✔ Inconsistencies and biases in raters, which oblige the development of very clear, elaborate and costly Scoring Rubrics. Performance assessments also require expensive training processes for raters, in order to ensure consistent scores across raters and occasions.
- ✔ Marking is costly and often too time-consuming, which makes formative use of the results difficult.
- ✔ The complexity of the tasks often gives rise to multidimensional structures –which hinder the use of unidimensional IRT models for estimation, equating or calibration – and to differential item functioning.
- ✔ More research is needed on differential item functioning in performance assessments compared to other formats, and on the influence of motivational factors.
- ✔ Although their apparent validity is clear, the different forms of evidence of psychometric validity need further study. Very important in this regard are certain irrelevant aspects on which raters focus, and whose influence must be removed. Substantive evidence related to processes should continue to be studied with measurement models that permit their assessment, and research should also focus on learning growth processes. Furthermore, it would be advantageous to explore in more depth the evidence of relations with other variables.

Current developments of psychometric models both in the field of IRT and in other frameworks (Mislevy, 2006), and which permit the involvement of components related to process, represent a significant advance. Multidimensional and hierarchical IRT models will also make it possible to deal with some of the limitations mentioned above. More research is need on the appropriate combination of tasks in multiple-choice or short-answer format and performance assessment tasks with a view to optimizing information.

The use of new technology can make it possible to address many limitations. Computerized presentation and response using adaptive tests allows considerable reduction in testing time, improving the representation of the domain. They also permit the use of dynamic tasks, such as those employed in patient diagnosis, as well as increasing authenticity through the inclusion of a range of resources (graphics, video, audio, reference materials, etc.). Furthermore, they improve the recording of processes through follow-up, highlighting evidence of substantive validity. The emission of responses via computer means that the interference of some irrelevant aspects related to writing and forms of expression can be avoided. The development of automated marking systems should continue, given the considerable potential benefits in relation to cost.

Finally, to the question of whether performance assessments should substitute traditional formats such as multiple-choice, the answer is no, since there are many aspects of assessment for which such formats, more economical in time and money, are quite adequate. The ideal approach is the appropriate combination of the different types.

In the present article we have offered a brief overview of performance assessments. Those interested in learning more about the topic can find an extensive treatment in the references cited by Johnson et al., (2009) on applications in education and in accreditations; the work by Thornton and Rupp (2006)deals quite extensively with the question of assessment centers. Furthermore, examples of performance assessment tasks typically used in educational assessment can be found among the items published by the PISA study (http://www.pisa.oecd.org), and information on performance assessments in certifications and accreditations is available on the websites of the American Board of Pediatric Dentistry (http://www.abdp.org/pamphlets/oral_handbook.pdfv, of the National Board of Medical Examiners

(http://www.usmle.org/Examinations/step2/step2ck_content.html) and of the previously-mentioned National Conference of Bar Examiners (http://www.ncbex.org/multistatetests/mbe).

REFERENCES

Abad, F.J., Ponsoda, V., & Revuelta, J. (2006). *Modelos politómicos de respuesta al ítem* [*Polytomous item response models*]. Madrid: La Muralla.

Ackerman, T.A., Gierl, M.J., & Walker, C.M. (2003). An NCME instructional module on using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*, 37-53.

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Arora, A., Foy, P., Martin, M.O., & Mullis, I.V.S. (Eds.) (2009). TIMSS *Advanced 2008: technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks, CA: Corwin Press.

Arthur, W., Day, E.A., NcNelly, T.L., & Edens, P.S. (2003). A meta-analysis of the criterion-related validity of assessment center dimension. *Personnel Psychology, 56*, 125-154.

Ayala, C.C., Shavelson, R.J., Yue, Y., & Schultz, S.E. (2002). Reasoning dimensions underlying science achievement: the case of performance assessment. *Educational Assessment, 8*, 101-121.

Bennett, R.E. (2004). *Moving the field forward: Some thoughts on validity and automated scoring* (ETS RM 04-01). Princeton, NJ: Educational Testing Service.

Bock, R.D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.

Brennan, R. L. (2000a). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. (2000b). Performance assessment from the perspective of the generalizability theory. *Applied Psychological Measurement, 24*, 339- 353.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33–63.

Brummel, B.J., Rupp, D.E., & Spain, S.M. (2009). Constructing parallel simulation exercises for assessment centers and other forms of behavioral assessment. *Personnel Psychology, 62*, 137-170.

Clauser, B. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24*(4), 310–324.

Cohen, A., & Wollack, J. (2006). Test administration, security, scoring and reporting. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 355-386). Westport, CT: American Council on Education/Praeger.

Cole, N.S., & Moss, P.A. (1989). Bias in test use. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 201-220).

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L., Linn, R., Brennan, R., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement 57*, 373–399.

DeMars, C. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*, 55- 77.

Dorans, N.J., Pommerich, M., & Holland, P.W. (Eds.). (2007). *Linking and aligning scores and scales*. New-York: Springer.

Drasgow, F., Luecht, R.M., & Bennett, R.E. (2006). Technology and testing. In R.L. Brennan (Ed.), *Educational Measurement,* 471-515.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 300–396.

Engelhard, G., Gordon, B., Walker, E.V., & Gabrielson, S. (1994). Writing tasks and gender: Influences on writing quality of black and white students. *Journal of Educational Research, 87*, 197-209.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

Fitzpatrick, R., Ercikan, K., Yen, W.M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11*, 95-208.

Fitzpatrick, R., & Morrison, E. (1971). Performance and product evaluation. In R. Thorndike (Ed.), *Educational measurement* (pp. 237–270). Washington, DC: American Council of Education.

Frederiksen, J.R., & Collins, A. (1989). A systems approach to

educational testing. *Educational Researcher, 18*, 27-32.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007a). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4-19.

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

Gorin, J.S., & Embretson, S.E. (2006). Item difficulty modelling of paragraph comprehension items. *Applied Psychological Measurement, 30*, 394-411.

Hambleton, R.K. (1996). Advances in assessment models, methods, and practices. In D.C. Berliner and R.C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 899-925). New York: Macmillan.

Hambleton, R.K. (2000). Advances in performance assessment methodology. *Applied Psychological Measurement, 24*, 291-293.

International Taskforce on Assessment Center Guidelines (2000). Guidelines and ethical considerations for assessment center operations. *Public Personnel Management, 29*, 315-331.

Johnson, R.L., Penny, J.A., & Gordon, B. (2009). *Assessing performance: designing, scoring, and validating performance tasks*. New York: Guilford Press.

Kane, M.T. (2004). Certification testing as an illustration of argument-based validation. Measurement: Interdisciplinary *Research and Perspectives, 2*, 13-170

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York: Springer.

Lance, C.E. (2008). Why assessment centers do not work the way they are supposed. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 84-97.

Lane, S., Parke, C.S., & Stone, C.A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment, 8*, 279-315.

Lane, S., & Stone, C.A. (2006). Performance assessment. In Brennan (Ed), *Educational Measurement*, (4th ed., pp. 387–431). Westport, CT: American Council on Education and Praeger.

Linn, R.L. (1993a). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*, 1-16.

Linn, R. L. (1993b). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83-102.

Linn, R.L., Baker, E. L., & Dunbar, S.B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Livingston, S.A. (2009). *Constructed-response test questions: Why we use them; how to score them* (R & D Connections, nº 11). Princeton, NJ: Educational Testing Service.

Livingston, S.A., & Rupp, S.L. (2004). *Performance of men and women on multiple-choice and constructed-response tests for beginning teachers*. (ETS Research Report No.RR.04-48). Princeton, NJ: Educational Testing Service.

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

Madaus, G., & O'Dwyer, L. (1999). A short history of performance assessment. *Phi Delta Kappan, 80*(9), 688–695.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–103). Washington, DC: American Council on Education.

Messick, S. (1994). The interplay of evidence and consequence in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. (1996). Validity of performance assessments. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington, DC: National Center for Education Statistics.

Miller, D.M., & Linn, R.L. (2000). Validation of performance assessments. *Applied Psychological Measurement, 24*, 367-378.

Mislevy, R.J. (2006). Cognitive psychology and educational assessment. In R.L. Brennan (Ed.), *Educational Measurement*, pp. 257-305.

Mislevy, R., Steinberg, L., & Almond, R. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496.

Mislevy, R., Wilson, M., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. In T. Kellaghan and D. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 489–532). Boston: Kluwer Academic.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

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

Muraki, E., Hombo, C.M., & Lee, Y.W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*, 325-33.

National Conference of Bar Examiners (NCBE) and American Bar Association (ABA). (2005). *Bar admission requirements*. Available at http://www.ncnex.org/tests.htm. ECD (2007). PISA 2006 *Science Competencies for Tomorrow's World*. Paris: OECD.

O'Neil, H.F., Subgure, E., & Baker, E.L. (1996). Effects of motivational interventions on the National Assessment of Educational Progress Mathematics Performance. *Educational Assessment, 3*, 135-157.

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation*, 13, nº 4. Available at http://pareonline.net/getvn.asp?v=13&n=4

Parke, C.S., Lane, S., & Stone, C.A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation, 12*, 239-269

Patz, R.J. (2006). Building NCLB science assessments: Psychometric and practical considerations. *Measurement, 4*, 199-239.

Penfield, R.D., & Lamm, T.C.M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practices, 19*, 5-15.

Prieto, G., & Delgado, A.R. (2010). Fiabilidad y validez [Reliability and validity]. *Papeles del Psicólogo, 31*(1).

Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25-36.

Reckase, M.D. (2009). *Multidimensional item response theory*. New York: Springer.

Rupp, D. E., Gibbons, A.M., Baldwin, A. M., Snyder, L. A., Spain, S. M., Woo, S. E., et al. (2006). An initial validation of developmental assessment centers as accurate assessments and effective training interventions. *Psychologist-Manager Journal, 9*, 171–200.

Rupp, D.E, Thornton, G.C., & Gibbons, A.M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology, 1*, 116-120.

Ryan, T. (2006). Performance assessment: Critics, criticism, and controversy. *International Journal of Testing, 6*(1), 97–104.

Ryan, J. M., & DeMark, S. (2002). Variation in achievement scores related to gender, item format and content area tested. In G. Tindal and T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: validity, technical adequacy, and implementations issues*, (pp. 67–88). Mahwah, NJ: Lawrence Erlbaum Associates.

Salgado, J.F., & Moscoso, S. (2008). Selección de personal en la empresa y las AAPP: de la visión tradicional a la visión estratégica [Personnel selection in industry and public administration: from the traditional view to the strategic view]. *Papeles del Psicólogo, 29*, 16-24. http://www.cop.es/papeles.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

Stecher, B., Klein, S., Solano-Flores, G., McCaffrey, D.M. Robyn, A., Shavelson, R., & col. (2000). The effects of content, format and inquiry level on science performance assessment scores. *Applied Measurement in Education, 13*, 139–160.

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Stiggins, R. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practices, 6*(3), 33–42.

Thornton, G.C., & Rupp, D.E. (2006). *Assessment centers in human resource management*. Mahwah, NJ: Erlbaum.

United States Medical Licensure Examination (2009). *Examinations*. Available at http://www.usmle.org/examinations/index.html.

Van der Vleuten, C., & Swanson, D. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*, 58-76.

Welch, C. (2006). Item and prompt development in performance testing. In S. Downing and T. Haladyna (Eds.), *Handbook of test development* (pp. 303–327). Mahwah, NJ: Erlbaum.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

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

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

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