Psicológica (2001), 22, 43-66.

The definition of achievement and the construction of tests for its measurement: A review of the main trends

Salvador Algarabel and Carmen Dasí*

Universitat de València, Spain

This is a review paper in which different definitions of achievement are analyzed and different possibilities for test construction are explored. A first characterization of achievement is accomplished through the analysis of construct representation. From this perspective, the behavioral approach focuses more on the end result, whereas the cognitive approach is more process centered. In a second stage, this review analyzes the data about nomothetic amplitude: the relationships between achievement and aptitudes, socioeconomic status, and changes over time. The final section offers a view of the possibilities and difficulties involved in the attempt to substitute traditional methods for performance assessment methods. Given the difficulties and cost in development time, scoring and other variables, the review concludes by assigning a major role to computer technology in assessment, if performance assessment is going to have a chance to achieve widespread use.

Keywords: achievement test, item response theory, automatic item generation, automatic scoring, performance assessment and test construction.

Test construction was originally driven by an interest in the measurement of mental abilities. Their conceptualization drove the technology that for a long time has been applied to test construction. When achievement began to be measured, the principles of test construction applied were identical to those used in the measurement of abilities (Glaser and Silver, 1994; Levine, 1976). Beginning in the sixties, the cognitive movement started to question the lines along which achievement had been previously defined and measured. Selection is often the main goal of ability

* This research was supported by a grant from the "Dirección General de Investigación Científica y Técnica (PB/97-1379)" from the Spanish Ministry of Education and Science.

testing, whereas diagnosis at an individual or institutional level, in addition to accountability, is also a goal of achievement testing. These differences and others have led to the development of new tools in test construction and to a divergence from the technology developed in aptitude measurement. In general, some of these advances are: development of criterion referenced tests, attempts to generalize the use of open response as an alternative to multiple choice item, and the increasing role within the educational movement of authentic assessment.

This is a paper where a current view of achievement is going to be established with reference to contemporary psychological and educational movements. To carry out this task, the distinction made by Embretson (1983) between construct representation and nomothetic amplitude is followed. First, we will analyze achievement very briefly from the point of view of the cognitive or behavioral mechanisms involved in solving individual items (construct representation). Next, we will look at the relationships between scores on the achievement test and other cognitive measures (nomothetic amplitude). Both aspects of the definition of achievement are important one way or the other in the specification or application of a test. Finally, We will attempt a critical review of the possibilities offered by the new measurement models to test construction in the assessment of achievement. Part of this final appraisal is the recognition of the impact of the widespread use of computers and the establishment of connectivity (Internet) as special engines that will drive future changes in achievement evaluation. The hope is that the new ideas can be correctly implemented with the help of the new technology, and as a result, the psychometric foundations of the new instruments may be well established. A final word about the use of the word *test* and *assessment*. In this paper, both terms are used interchangeably, although test is a more restricted term than assessment, indicating a more standardized and quantitative approach to measurement.

Achievement from the point of view of construct representation

In the Standards for test construction (APA, 1999) achievement is viewed basically as the competence a person have in a area of content. This competence is the result of many intellectual and nonintellectual variables, although in this paper we concentrate exclusively on the former.

The scientific study of achievement encompasses data coming from experiments with word lists as well as from the acquisition of complex domains, like computer programming, mathematics, or the way in which people solve physics problems. At the experimental level, achievement is referred to as acquisition, learning, or knowledge representation, sometimes

depending on theoretical biases. Achievement is the word preferred in the educational or psychometrics fields, being sometimes characterized by the degree of inference required on the part of the student to give a response, and by the type of reference to a cognitive process made explicit in the measurement tool. As we said earlier, we are going to sketch some of the theoretical positions on achievement that have influenced the way in which it is measured, in order to later introduce the problems involved in test construction.

In the 1950's and early 60's usually students had to master the "basic facts" (e. g. Schoenfeld, 1992), meaning the reproduction of declarative knowledge. It was thought that these basic facts were necessary to build further abstract rules, and little reference was made to possible cognitive processes, no matter what complexity of inference was required from the student. Although it is true that without basic facts there is little possibility for abstract reasoning, the influence of behaviorism made it unacceptable to refer theoretically to these abstract processes in the way we do today. An extreme historical view of this approach sought to analyze an achievement field and establish a small, step by step progression of knowledge with the goal of letting the student master the domain (e.g. Holland and Skinner, 1961). The approach was called "programmed instruction". Chaining, associations, interference or transfer were common analytical tools for this approach in the early 60's in the study of learning at the experimental as well as the educational levels.

Cognitive psychology produced a shift from the study of behaviour to its unobservable psychological antecedents. The cognitive analysis of achievement means to get into the experimental study of memory storage and retrieval. From the cognitive point of view, achievement must be a construct that should refer to the different stages of knowledge acquisition. The end product; that is, the knowledge that characterizes the expert, is a highly structured set of mental models built after long sessions of practice. The consequences are that the expert can bring into play sophisticated strategies and take into account large bodies of knowledge without the usual working memory limitations. The studies and most accepted model on short term memory (Baddeley, 1986; Baddeley and Hitch, 1974) and data on memory span changes (Chase and Ericsson, 1981) clearly indicate that this system plays a crucial role in knowledge acquisition and reasoning. The amount of information processed by the system is always limited to a reduced number at least they are chunked. When a subject is faced with a reasoning task he has to integrate background and external knowledge, consuming limited resources. When the information is completely new and

of a very abstract nature, then the limitations of the system are at its maximum.

Work on experts, in such diverse fields as Physics or chess (Anzai, 1991; Charness, 1991; Ericsson, 1996; Ericsson and Smith, 1991), show that the expert is characterized by a well organized abstract body of knowledge based on general principles as well as specific knowledge related to the field of expertise. The amount of practice required to become an expert leads to very structured and compact schemas that will allow bypassing the working memory limitations. As part of this knowledge, the expert also has a set of general and specific metacognitive strategies for dealing with particular problems to be solved. These strategies can take into account more and more information, given the highly structured nature of the long-term memory. Educationally, achievement may be defined (Niemi, 1999) as the mastering of major concepts and principles, important facts and propositions, skills, strategic knowledge and integration of knowledge. More systematically, achievement is sometimes fractionated into knowledge components (Ruiz-Primo, 1998), like declarative, procedural and strategic. The declarative knowledge is composed of domain specific content, whereas the procedural and strategic refer to specific production systems (Anderson and Lebiere, 1998) and specific heuristics (Schoenfeld, 1992). The cognitive system has also the ability to monitor the process and use nonspecific strategies that are also a part of our proficiency in achievement. These different components of achievement develop conjointly and cannot be treated separately.

In summary, achievement is the competence of a person in relation to a domain of knowledge. What we can externally observe is performance. The current view states that to reach a specific level of performance it may be necessary to bring into play complex cognitive tools like strategies, heuristics or skills. No doubt that the end result and the type of means to reach it must be correlated (e. g. Willson, 1989), a fact often overlooked. A difficult problem can only be solved after a well organized body of knowledge is consulted and the appropriate metacognitive skills are used to reach a solution. The question then is what can be gained or lost, when taking into account the whole process, as when an open response is assessed, or just the final solution, as in multiple choice. From the point of view of measurement instrument, one can argue that if there is no compromise in reliability; that is, if the evaluation of the whole open response is carried out with a high level of precision, the measurement of the open response will increase validity. However, a more critical point has to do with the consequences of focusing, from an educational perspective, on the cognitive processes supposedly involved in the final performance. If

the cognitive processes that lead to expert performance must be taken into account, the definition of achievement from a complex cognitive view has long reaching consequences, because by emphasizing these aspects we are promoting a level of expertise not reachable by other means. This is the position of most proponents of the new educational movements which try to reform the testing procedure.

Achievement from the point of view of nomothetic amplitude

In the previous section we have discussed achievement from the point of view of its internal cognitive mechanisms; now we turn to the relationship of achievement and alternative constructs. Most of these relationships are malleable (achievement and family atmosphere, socioeconomic status, country, ethnicity) and others are more fixed (achievement and aptitude). The analysis of the changeable influences, in the case of achievement, serves the purpose of diagnosing educational or schooling programs, and making decisions on the basis of the data. The purpose of examining the relationship between achievement and aptitude is uncertain.

There are a lot of metanalytic studies to illustrate such relationships, for example, in family influence (Lytton, 2000), in gender differences (Nowell & Hedges, 1998; Stumpf & Stanley, 1996), in social factors (Ma & Kishor, 1997), and so ong. An exhaustive review of them will exceed the limits of the present article. To illustrate this point some studies will be concisely described next. They serve as example of the main outcomes of such reviews.

If we correlate an aptitude test (for example the Raven) with an achievement test (like any scale of the NAEP), values in the neighborhood of .50 or higher (see Neisser et al., 1996) are obtained, and when we correlate aptitude tests with school performance, correlations are also high. The performance assessment movement believes that this high correlation may be due to the forced normative nature of many achievement tests. If the test is intended to discriminate among people, the achievement tests are forced to become more and more like aptitude tests (Glaser and Silver, 1994; Levine, 1976). This is probably an extreme view and it seems indisputable that the more we define achievement as inferential, the higher the relationship between aptitude and achievement is expected to be. As an example, recent studies (Pasnak, Willson, and Whitten, 1998) show high correlations between the Peabody Individual Intelligence, and the Peabody Picture vocabulary tests. High correlations are also found with Piagetian tasks in populations of children with mild retardation (see also Edwards and

Kirby, 1964; Throne, Kaspar and Schulman, 1965, or Coleman and Cureton, 1954).

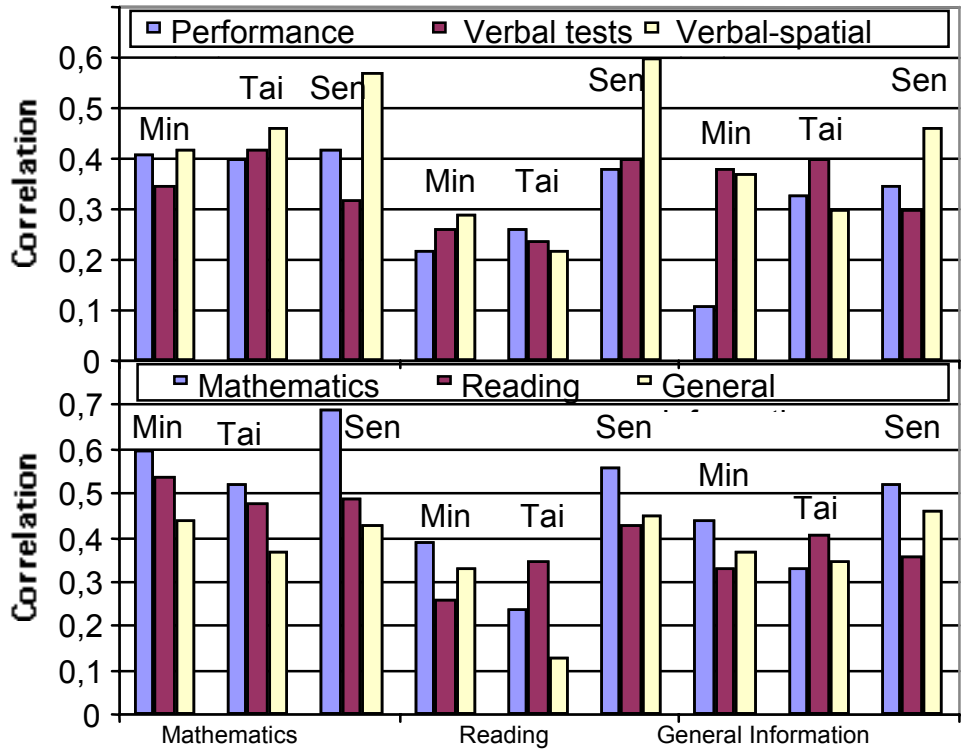

Figure 1. On the top, correlations among cognitive measures taken at grade 1 (performance, verbal and verbal-spatial) and achievement tests taken at grade 11 (mathematics, reading, and general information). Onto the bottomleft, interintercorrelations among achievement tests (mathematics, reading and general information) given at grade 1 and 11 as a function of three different populations (Minnesota, USA; Taipei, Taiwan; Sendai, Japan) are shown. To the right, intercorrelations among cognitive measures taken at grade 1 (performance, verbal and verbal-spatial) and achievement tests taken at 11 (mathematics, reading, and general information). The graph has been drawn from the data presented by Chen, Lee and Stevenson (1996, p. 755).

Figure 1 is the representation of part of the results of a large-scale study (Chen, Lee, and Stevenson, 1996) in which the relationships among several different measures of achievement and cognitive abilities are examined for three different populations: Minnesota (USA), Taipei

(Taiwan) and Sendai (Japan). The investigation takes achievement into account at different grades (grade 1 and 11) in mathematics (computation, arithmetic, algebra, etc.), reading and general information, and examines the relation with some cognitive measures. Nine culturally appropriate cognitive tests were constructed and they were combined into three summary scores: performance tests (i.e., coding, spatial relations, auditory memory, and perceptual speed), verbal tests (i.e., verbal memory, vocabulary, serial memory for numbers, and serial memory for words), and verbal-spatial test (i.e., verbal-spatial representations). All of the correlations in this figure (excepting the two lowest) reached statistical significance.

On the top, the graph shows correlations between cognitive tests given at grade 1 and achievement tests given at grade 11, particularly high between the verbal-spatial test and the most of the achievement measures and populations. On the bottom, the intercorrelation pattern among achievement tests at grade 1 and 11 for mathematics, reading and general information shows a consistent positive trend for the three countries. Early achievement tests can predict later measurements, particularly in fields like mathematics. The type of measures used in this study range from multiple choice, predominantly memoristic, to traditional mathematics problems with open responses. This study also shows that achievement for Japanese and Chinese people is higher than for North-American people. However, concerning the comparison among countries, there is a bigger and more complete research carried out to study performance in mathematics. We are referring to the Third International Mathematics and Science Study (TIMSS) whose last report is from 1995-96. This study also shows that Japanese and South Korean children show, in general, superior performance to children of a large series of countries in content areas like algebra, geometry, measurement and proportionality (U.S. Department of Education, 1996).

Achievement differs also among different socioeconomic status groups. White (1982) showed positive correlations between academic achievement and some measures of social status like family income, parent's occupation, etc. In some of the studies, White took into account the student as the unit of analysis, and in others, he used aggregated units of analysis, such as the school or the district, for computing the correlation. The study showed that the correlation is "only" of about .22 when it is computed over individual subjects, and goes up to about .70 for the aggregated units. Probably there are other variables moderating the relationship and not social status per se. The author of the study is willing to conclude that familiar atmosphere is the variable of interest.

In a study William and Ceci (1997) found differences between American people of African (black) and of European origin (white) in mathematics and reading (Figure 2).

People was tested when they were 9 and 13 years old. Although the difference has been narrowing in Mathematics, in Reading it seems to have stabilized in recent years, showing no sign of disappearance.

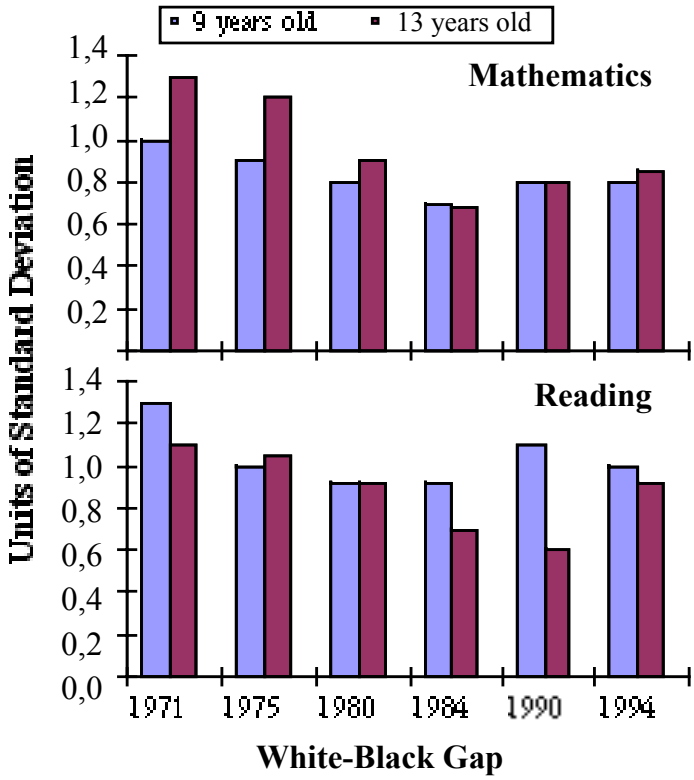

Figure 2. Differences between people of African and European origin on achievement in mathematics and reading (source William and Ceci, 1997, p. 1229).

In conclusion, the achievement construct is related, at least moderately, to aptitudes as defined by psychometric tests. Other indicators, like the influence of socioeconomic status and the rest of the reviewed variables, may help to implement policies to equilibrate undesired influences on achievement.

In the forthcoming sections we confront the measurement of the complex construct that we have previously defined. Overall, we have to measure a heterogeneous variable, though more or less easily decomposable

into components, the isolated proficiencies on those components are highly interrelated. We should not lose sight of the fact that a high level of declarative knowledge accompanies the presence of a high level of proficiency on procedural and strategic knowledge.

Although performance assessment has had a big impact on educational researchers, the consequences for achievement testing are not always as positive as it might be thought. The use of complex assessment tasks, although trying to solve the traditional problems implied in the use of multiple choice items, introduces some additional ones (e. g. Wainer and Thissen, 1994). A non exhaustive list would include problems with low reliability (Shapley and Bush, 1999), presence of face but dubious construct validity, increases in cost and testing time (Luckhele, Thissen and Wainer, 1994), and lack of fairness in certain testing situations (Wainer and Thissen, 1994). On the other hand, some studies show the dissatisfaction produced in teachers, particularly science teachers (Baker, 1998), by the use of some of the performance assessment instruments like the portfolio, not to mention the lack of data supporting the psychometric characteristics of its different assessment procedures (e. g. Shapley and Bush, 1999; Stecher, Barron, Kaganoff and Goodwin, 1998; Strong, and Sexton, 1996, 1997). The conclusion that we can extract is that if new forms of assessment are to be used, there is a need to free teachers, and sometimes students, of the burden in development and scoring time imposed by these procedures. This is the place where computers may play a very important role, as will be pointed out later on. Item types, and scaling models are the hottest topics with regard to the different trends that achievement testing may follows in the future. But the use of networking computers is the main ingredient necessary for the implementation of any theoretical and methodological advance.

The Measurement of Achievement: Item type, scoring, and reference

Having seen the different approaches to the definition of achievement, one is left with the task of surveying the different trends that assessment of achievement may follow in the future. By new perspectives we are referring to the survey of possibilities opened up by the union of a cognitive complex definition of the construct intended to be measured, and the technological advances introduced by modern computers. Tests and assessment instruments used today can be characterized by two shifts. In the first place by a validity shift from behavioral to cognitive, and in the second, by a psychometric shift from the application of the linear to item response models. The next generation of tests, what Bennett (1998a) calls generation

"R", must introduce important theoretical changes. These changes will be materialized by new item and response formats that will form the assessment environment, and that initially may be separated from the learning environment. In a later stage, learning and assessment will merge and proceed conjointly into the electronic environment set up by Internet. In the next sections, we analyze the new forms of assessment. Quickly it will become apparent that if the new forms of assessment survive in the future, this survival will come from the generalized use of computers, making possible automatic item generation, the construction of high quality multimedia items and the automatic scoring of open responses. Finally, the generalized used of Internet will dramatically change the educational rules.

The different approaches to the definition of achievement may differ in at least three steps with regard to test construction. The first is the type of item necessary to measure achievement. The second is the kind of reference that should be built with regard to score interpretation. And the third is the types of interpretations for the scores and, in consequence, the kind of validity that must be estimated for the instrument.

With regard to the type of item, achievement can be examined along a continuum, which extends from multiple choice to performance assessment tasks. The issue of item type needs to be analyzed from the point of view of reliability, validity and standardization. Traditionally, the classic psychometric approach pointed towards the multiple-choice item as a good way to measure achievement in an economical and standardized way. In fact, the multiple choice item is one of the three characteristics defining what van der Linden (1986) calls the “classical complex” approach to measurement. That is, the multiple choice item, although simple, was supposed to be able to measure knowledge regardless of its complexity. Given that the multiple-choice item was valid, the test was assembled from a set of independent items to preserve the conditional independence assumption required by the measurement models. The student true score is his/her expected response throughout the whole domain, which is a function of the accumulated sum of the items sampling the domain (Mislevy, 1996).

With the new educational movements, the multiple choice item has been at the center of criticism of the traditional approach to test construction. If achievement was defined as in an earlier section, the multiple-choice item was argued to lack validity to capture the construct and, additionally, has very undesirable consequences on the educational system (consequential validity). The natural way out has been to recur to the use of open responses, or more "real" tasks like those involved in authentic or performance assessment. Johnsen and Ryser (1997) says that "I encourage the improvement of procedures to measure educational

performance; I emphasize the importance of alternatives to multiple choice to promote in the student the mastering of comprehension more than recognition of correct alternatives".

As the cognitive movement has stressed the definition of achievement from the point of view of high order cognitive processes, its proponents have turned towards complex forms of assessment. In fields where people need not to be scaled, like in basic human learning and memory research, structural knowledge is sometimes assessed by such spatial methods as multidimensional scaling. In educational settings, performance assessment is today the expression to assess achievement. The performance assessment movement has remained largely qualitative and diverse, and has generally been conceptualized as a type of criterion-referenced measurement. The proposed assessment ranges from tasks that require short elaborate answers, to class projects developing over a more extended period of time, on demand tasks over which the student has little control, or portfolios, where the student collects pieces of his work over time (e. g. Baker, 1998). A common denominator of all these tasks is the use of complex types of items that require the production of open responses.

However, although the new types of items are supposedly more valid they are also less reliable, not to say about their more costly application and scoring. A balanced view should defends the position that there is no single or unique item type able to be defended as the best way of assessing knowledge. The hope is that modern technology helps to better standardized the new and more complex types of items removing their main criticism. This is a way already started with computerized testing (e. g. those tests administered by the Educational Testing Service), using items that are a mixture of open response and multiple choice in a way that tries to maintain a high level of standardization and provide an increment of construct validity. Another possibility is to sophisticate the multiple choice item. One of these sophistications is the testlet (Wainer and Kiely, 1987, p. 190) defined as "group of related items in a unique area of knowledge which behave like a unit and that includes a predetermined number of steps that the examinee may follow". The testlet has been used in applications of item response theory and where a common scenario is defined for a series of items that by their nature violate the necessary assumption of conditional independence required by all scaling models. So one important trend for future test development is the increasing use of complex item types approaching traditional characteristics by virtue of technological advances.

Given all these trends, the use of computers is deemed completely necessary to fulfilled the requirements of modern educational movements. Three aspects of this influence are especially relevant: automatic item

creation, the implementation of multimedia items of high quality, and finally, the automatic scoring of responses. These three developments are not new, but computers have given new possibilities to their implementation. However, their significance for achievement testing should fully be appreciated, taking into account the new educational movements. Performance assessment is an example. According to some of its proponents (e. g. Baker, 1998), the assessment of higher order thinking requires the use of complex tasks (see also Bejar, 1998). We have already discussed some of the pros and cons of this proposition. Suffice is to say (Baker, 1998) that assessment based on complex models of learning means that the tasks may be difficult and expensive to develop. The point of view here is that performance assessment may generalize to large testing or stand alone applications if computers take care of some of the problems that it generates.

Take, as an example, the definition of problem solving used at CRESST by O'Neil, Jr. and Schacter (1997). According to this model, problem solving implies the assessment of four domains: content understanding, problem solving strategies, metacognition and motivation. If we were measuring this construct of problem solving by traditional means, the final tool would be impossible to implement due to cost and time considerations, not to mention psychometric objections. These, and some other objections imposed by the very nature of performance assessment, preclude the use of traditional assessment instruments, and the needed improvement of the performance assessment instruments from its current state to a more healthy state. Only if the computer takes care of these problematic aspects, will performance assessment gain credibility. The first aspect where computers may help is in the presentation of complex tasks, like high quality multimedia items (Bennett, 1998a). The point is that the possibility of presenting multimedia items gives place to the measurement of complex aspects of the construct under measurement because now it will be possible to vary systematically the context of the application of knowledge in a practical way.

However, when the construct has to be measured comprehensively in a short period of time, like in traditional adaptive testing, we need to be able to generate a large number of items with known psychometric properties, in advance. If this process had to be made manually, cost ant time considerations would preclude its full implementation. For these two main reasons automatic item generation is a very important line of future development. We understand automatic item generation here to mean the results of the implementation of an algorithm that will be able to generate item exemplars from a common schema (other possibilities in Revuelta and

Ponsoda, 1999). One prototype of a modern item generator is the Math Test Creation Assistant (MTCA; Bejar, 1998; Bennett, 1998a; Singley and Bennett, 1998). The MTCA applies schema theory to generate items characterized in terms of a set of equations to produce variants implementing a particular set of equations more restricted than the original schema set. A second step of this program is the ability to predict the item difficulty given a preexisting known relationship between item characteristics and psychometric statistics. With this tool, the developer may produce a family of models according to variations and commonalities among equations and variables. The models may be hierarchically organized, and when instantiated, lead to the items' variants.

However, the most important point to be taken into account, in our opinion, is the possibility of scoring complex types of responses. We have already indicated that the multiple choice item is at the center of the criticism by the performance assessment movement. It is thought that multiple choice test only measures factual knowledge (e. g. Glaser and Silver, 1994; Sugrue, 1996; Shilpi,, 1995). However, the multiple choice item was adopted at a time when no other possibilities existed to maintain standardization and keep cost and scoring time low. Early or more basic attempts to automatically score complex responses (Martinez and Bennett, 1992) are a basic evolution of the multiple-choice item, like the grid-in type. Present day computers make it possible to widen the scope of this automated scoring through the use of sophisticated presentations or the use of complex parsers to analyze a less restricted response. In fact, it would be fair to indicate that the main reason for defend the use of multiple-choice items is the standardization that it allows. Present day computers have changed this perspective in many fields, particularly in mathematics, due to the sophisticated possibilities of choice that allow a restricted open response to be made.

Automatic scoring can also be applied to multiple line responses, like the typical process answer given to a mathematics problem. Schema theory (Singley and Bennett, 1998a) may be used in these cases. It is, however, in the domain of natural language where the scoring of responses is most necessary. There are multiple solutions to this problem. The approach taken in the Educational Testing Service has been (e. g. Bejar, 1998; Burstein, Kukich, Wolff, Lu, and Chodorow , 1998) to decompose the response into a set of features and subfeatures that finally can be aggregated to a unique score (e. g., Bejar and Bennett, 1999; see also Bennett, 1998b). The result of this approach is an experimental system called e-rater (e. g. Burstein, Kukich, Wolff, Lu, and Chodorow , 1998) whose operational way of work in grading an open response is to let the computer use the same

characteristics followed by a human rater when grading. Just to give an example, in the case of grading the Graduate Management Admission Test (GMAT), the computer extracts syntactic, rhetorical and topical content features, according to computational linguistics techniques. The computer is given training with a series of human graded responses, allowing it to establish values for the previous features, and establishing finally a score based on stepwise regression. The data show a very high correlation between computer and human in a variety of situations.

There are other commercial procedures for scoring open responses (e. g., INQUIZIT, or INTELLIMETRIC http://www.intellimetric.com/) that we are not going to analyze here in detail, except for a few words about the latent semantic analysis model (LSA, e. g. Foltz, Kintsch and Landauer, 1998; Landauer and Dumais, 1997 http://LSA.colorado.edu/). LSA is a long term memory model that uses singular vector decomposition to generate a knowledge map from a set of, say, text inputs to the algorithm. One of the most noteworthy characteristics is that this "map" makes no reference to any "semantics" of the text. On the contrary, it generates a dimensional space of "concepts" by "context". The context may be a sentence, a whole paragraph or a whole text, and the dimensional space may be as high as 200 or 300 dimensions or more. Preliminary studies indicate that the reliability with which the computer scores the response of a student, taking as reference the "mental model" developed by the LSA procedure, may be as high as that obtained by human raters. The relevant aspect of the LSA model is that it is at the same time a model of human long term memory (Landauer and Dumais, 1997).

With regard to the reference of the scores, if we take a view of achievement as similar to ability (Glaser and Silver, 1994) and see the instrument as normative, then item analysis will lead to the selection of the most discriminant item in relation to the population of subjects. Traditional indexes used for item analysis, like discrimination and difficulty, either by item response theory or by linear model estimations, will play a major role in the final composition of the test.

For a cognitive theorist it is more natural to define an achievement test as criterion referenced. With a criterion referenced test the emphasis shifts from prediction to diagnosis, and the student is assessed in absolute terms against a set of values previously set in the standards. The procedure for setting standards is a problem in itself (e. g. Berk, 1996; Cizek, 1993; Linn, 1998), although this is not the place to review it in detail. The important question is that, although discrimination must not be abandoned, recognition of the need to refer to the educational goals must be established. This current trend is not new (see Glaser, 1994) although its widespread

recognition has taken place only recently. Both aspects: norms and criteria are important, and most of modern educational tests are criterion and norm referenced.

The issue of the use and interpretation of scores (validity) takes us to two different positions. One is to use the achievement scores to predict other behaviors, in line with its use in selection settings (Glaser and Silver, 1994). In this case, predictive validity will be one of the main indexes of the test. Others may see the test as a way of reflecting the student evolution through several competence stages (e. g. Glaser, Lesgold and Lajoie, 1985; see examples in Mislevy, 1993, 1996), and content validity as an especially important index to be taken into account. In addition to its predictive value, the test must serve as a diagnostic tool and guide the explorations of the problems and stages of development of the cognitive system. A complete listing of the functions of assessment according to performance assessment theorists (Baker, 1999; Linn, 1999) are institutional accountability, program evaluation, instructional improvement, system monitoring, certification, selection, placement and fostering learning. Consequences, alignment of the instrument with established standards, fairness, transfer and generalizability, content quality, cognitive complexity, content coverage, linguistic appropriateness, meaningfulness, practicality and cost are the qualities that must be taken into account in a performance assessment instrument (Herman, 1999).

The measurement of achievement: Scaling models

In general, the different alternative models used for scaling the scores, are for the most part neutral in relation to the different aspects of achievement one likes to measure, although some scaling models offer more possibilities than others. This is the point where the different definitions of achievement and the techniques for its scaling must meet.

Until the late 70's, test construction for abilities as well as for achievement was carried out by applying the linear model (e.g. Lord and Novick, 1968). The model proposes a decomposition of an individual observed score in a true and error component, and statistically and conceptually draws from variance component analysis. This approach is still developing in what it is known as generalizability theory (Brennan, 1993; Cronbach, Linn, Brennan, and Haertel, 1977) and it is an instrument eagerly followed by people in applied fields given its facility to estimate classical reliability (unconditional) and other sources of error. It is well known that any test behaves better at certain ability ranges that at others, according to a variety of parameters. Therefore, a unique reliability coefficient is not a uniform indicator of precision. Nonetheless, given the central role played by

these indexes, a serious effort has been carried out (Brennan, 1996; Brennan and Won-Chang, 1997) to offer a systematic series of conditional reliability indexes to match some of the advantages of item response theory in this regard.

The linear model has been mostly applied to multiple choice items, where, by using traditional indexes like difficulty (delta or proportion correct), discrimination (biserial or point biserial) and normal distribution assumptions, item analysis was carried out, and the final scale was assumed to be of an interval nature. This is not the place to review in depth the problems that this scaling model has. Problems with the definition of parallel forms; lack of invariance of subject or item estimations, and in general, lack of axiomatic justification for the interval scale assumption are just a few. Particularly troublesome is the lack of invariance; that is, the dependency of item and subject ability estimations on the sample of subjects or items, respectively, used in calibration. For these reasons the linear model is not undertaking further developments at the hands of more mathematically oriented psychometricians (e. g. Ramsay & Stout, 1997).

On the other hand, since the 1980's, the model of choice for scaling has been provided by the item response theory (e. g. Lord, 1980). This theory assumes that the probability of correctly responding to an item is a nonlinear function of some type (logistic or normal distribution functions) of the subject ability and item difficulty. Depending on the specific model, this probability of responding may also be a function of the item discrimination and guessing, leading to the basic response models: one, two or three parameters. Item response theory has exploded with many new models (see van der Linden, and Hambleton, 1997) that extend the basic specifications and that can be easily applied to a variety of situations of importance in education. One of those is partial credit scoring. These extensions follow two very different paths. On the one hand, the varieties of models (e. g. Masters and Wright, 1997; Muraki, 1997; Samejima, 1997) that extends the basic latent trait model. In this case, the traditional assumption on conditional independence among the responses of a subject is achieved once a single latent variable, θ , is known and partial out. And, second, a family of multidimensional models has been developed (e. g. Embretson, 1997; Fischer and Seliger, 1997) that assumes explicitly the existence of a vector of abilities underlying a single response. It is not clear yet if this added complexity will mean a significant increase in modeling precision. Following this path, Pirolli and Wilson (1998) have presented a theory of measurement of the different aspects of learning based on the

multidimensional random coefficients multinomial logit model. These developments are indicators of the interest of theorists to apply item response theory to a very traditional field of theorizing like Experimental Psychology.

These item response theory extensions, like politomous and multidimensional models fit well with a view of achievement as a complex cognitive entity. In fact, we think that those extensions may put together the traditional standardized view of a testing instrument with the definition of achievement brought about by the performance assessment movement.. It is customary among performance assessment proponents the use of qualitative cognitive maps as the tool to measure complex knowledge acquisition. However, politomous response models (Algarabel and Ruiz, in preparation) like the generalized partial credit model (Muraki and Bock, 1996) can be used to scale relational knowledge in a more rigorous way than simple cognitive maps. The student is asked about the strength of the relations among a series of concepts, and instead of comparing it with the expert´s response and score it qualitatively, the judgment is analyzed according to an item response model.

The developments described so far may be impressive, but fall short of the changes that can be foreseen in a relatively near future due to the social and educational changes introduced by the global electronic networking. Internet will be priming more and more distance learning in very complex, interactive and cooperative environments. The student, by itself or in a group, will have the opportunity to integrate instantaneously many different sources of information of high quality in a social but physically distant environment. Now, the development of high quality environments (examples for medical training http://medicus.marshall.edu/) will be available to everyone, everywhere, instantaneously. One of the important consequences for the new forms of assessment is the integration learning and testing in the same process. This is one of the goals of the performance assessment movement and a necessity given the complexity of construct definitions and the need to look for reliability in the assessment process. The present empirical data (Wainer and Thissen, 1993) show that complex situations require much more testing time and length to equate the reliability of a multiple-choice equivalent test. If we were to use traditional testing tools to assess new construct definitions the enterprise would remain impossible. Internet provides new possibilities along this line because the student can be continuously monitored, and complex types of cognitive abilities, like strategies and heuristics, can be more easily assessed. This is the time for distant exams or web testing (Baker, 1999). The new extensions of item response models may play a role here. Particularly,

multidimensional models like the multidimensional random coefficients multinomial logit model as applied to the analysis of achievement (Pirolli and Wilson, 1998) may be used to get a variety of scores from the work of the student. In any case, in a continuously interactive environment, many different cognitive abilities must be simultaneously assessed from the same set of data, and decisions made with respect to the students´ training. So, different latent classes may assess different types of strategies whereas basic knowledge may be evaluated by a continuous variable (see a general framework in Mislevy and Verhelst, 1990).

CONCLUSIONS
THE FUTURE OF ACHIEVEMENT TESTING

In this paper, a review of the definition of achievement and of the trends for its measurement are presented. From the point of view of construct representation, achievement can be defined either as the resultant performance before the appropriate item or as performance in relation to a set of hypothetical cognitive processes. Research today is overwhelmingly cognitive, so instead of reviewing data supporting either position, we have preferred to establish the dichotomy in terms of what we can gain by measuring just the final correct response, or stratifying the domain as a function of the implied processes. To answer this question one has to establish the fact that if we implemented both approaches in relation to a common domain of knowledge, the scores would be correlated. That is, if the appropriate cognitive processes are not brought into play, one can not respond correctly. However, if a correct cognitive model were established, the advantage of this approach would be a considerable increase in precision, and secondly, have important consequences for the educational system. By emphasizing training on cognitive processes, expertise can properly develop, and the principle of "what you test is what you get" may become reality.

A side effect of the previous point is the type of item convenient to measure achievement. This is up to certain point an artificial question, given the present day technological advances. The motivations for the generalized use of the multiple choice item were standardization, and low economic and time costs. The advances in computer applications allow today the use of restricted open response (like items developed in the Educational Testing Service) or direct automatic scoring of open responses, in such a way that the old advantages of multiple choice can be considered extended to other item formats. Therefore, the only consideration for choosing one type of item over others, when computer testing is possible, must be the best way of tapping the construct. There is no compulsory need to use multiple choice

for the standardization it offers because this characteristic can be extended to other types of items due to technological advances.

Two developments in the context of educational measurement have taken place. On the one hand, a more psychometric approach has been designing new types of items with automatic scoring of complex responses and using item response theory in the context of adaptive testing. On the other, the performance assessment movement, with a mostly qualitative approach, has been using complex tasks and, with exceptions, not using sophisticated measurement models to assign scores. Before the new technological advances, the psychometric approach was criticized for lack of validity, whereas the psychometric properties of the performance assessment instruments leave much to be desired. The technological advances are going to change all of this in the near future. Given the complexity involved in performance assessment, only if computers take care of the most complex operations involved in the process and allow a continuous way of gathering data from the student, will the new assessment procedures survive. Of course, this survival involves the demonstration that the new instruments have greater validity and at least a comparable reliability to the instruments they are supposed to substitute. And at this point, item reponse theory, in the form of adaptive testing, remains unchallenged as the best scaling model to assign scores to people. The extensions of item response theory (van der Linde and Hambleton, 1997) may provide ways, in the near future, to assign scores in the way that many theorists think appropriate: Rash measurement, partial scoring, multidimensionality.

In conclusion, the definition of achievement can lead us more towards selection than to assessment of educational results, predictive validity than content validity, norm or criterion referenced, closed versus open response, although the alternatives are not completely disjointed. The cognitive movement claims that achievement is something more than "end facts". In consonance, many of their proponents try to evaluate these additional abilities by new forms of assessment, whose psychometric properties are yet to be well established. Some recent studies reveal that there are inconsistencies between the data obtained from performance assessment in relation to more traditional instruments like NAEP (Herman, 1999). However, item response theory may solve some of these inconsistencies, and the generalized use of computers may help to assess the complex definition of achievement defended by the educational theorists.

RESUMEN

La definición del rendimiento y la construcción de tests para su medida: una revisión de las principales tendencias. En esta revisión se analizan diferentes definiciones de rendimiento y se exploran posibilidades en la construcción de tests para su medida. Una primera caracterización del rendimiento se consigue a través del análisis de la representación del constructo. Desde esta perspectiva, la aproximación conductual, se centra más en el resultado final, mientras el enfoque cognitivo se centra más en el proceso. En segundo lugar, esta revisión analiza los datos sobre amplitud nomotética: relación entre rendimiento y aptitudes, *status* socioeconómico y cambios en el tiempo. La sección final ofrece una visión de las posibilidades y dificultades implicadas en el intento de sustituir los métodos tradicionalmente utilizados en la evaluación del rendimiento. Dada su dificultad y coste en términos del tiempo necesario para desarrollarlos, puntuarlos y otras variables, se concluye atribuyendo un peso mayor a las aplicaciones informáticas en evaluación, para que la evaluación conductual pueda tener mayor difusión.

Palabras clave: test de rendimiento, teoría de respuesta al ítem, generación automática de items, puntuación automática, evaluación del desempeño y construcción de tests.

REFERENCES

Algarabel, S. and Ruiz. J. C. (in preparation). *The measurement of knowledge structure by a generalized partial credit model*. Manuscript in preparation.

Anderson, J. R., and Lebiere, Ch. (1998). *The atomic components of thought*. N. J.: Erlbaum

Anzai, Y. (1991). Learning and use of representations for physics expertise. In K. A. Ericsson and J. Smith, eds., *Toward a general theory of expertise*. N. Y.: Cambridge University Press, 64-92.

American Psychological Association, American Research Association and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, D. C.

Baddeley, A. D. (1986). *Working Memory*. Oxford: Oxford University Press.

Baddeley, A. D. and Hitch, G. J. (1974). Working memory. In G. Bower, ed., *The Psychology of Learning and Motivation*, vol. 8, pp 47-90. New York: Academic Press.

Baker, E. L. (1998). *Model-Based performance assessment* (Technical report 465). Center for the Study of Evaluation Graduate School of Education and Information Sciences, Los Angeles: Ca.

Baker, E. L. (1999, February). *What´s the evidence of technology payoff*? . Paper presented at The National Conference, American Association of School Administrators.

Bejar, I. I. (1998, September). *Accelerating the use of technology in the classroom through assessment*. Paper presented at CRESST, Los Angeles, Cal.

Bejar, I. I., and Bennett, R. E., (1999). La puntuación de respuestas como un parámetro del diseño de exámenes: Implicaciones en la validez. In J. Olea, V. Ponsoda, and G.

Prieto, eds. *Tests informatizados. Fundamentos y aplicaciones*. Madrid: Pirámide. Pp. 53-59.

Bennett, R. E. (1998a). *Reinventing Assessment* (Technical Report). Princeton, N. J: Educational Testing Service.

Bennett, R. E. (1998b, September). *Using technology to improve assessment*. Paper presented at conference CRESST, Los Angeles, Cal.

Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education, 9*, 215-235.

Brennan, R. (1993). *Elements of Generalizability theory*. Iowa City, Ia: American College Testing.

Brennan, R. L. (1996). *Conditional standard errors of measurement in generalizability theory*. Iowa City: Iowa Testing Programs. The University of Iowa.

Brennan, R. L., and Won-Chan, L. (1997). *Conditional standard errors of measurement for scale scores using binomial and compund binomial assumptions*. Iowa City: Iowa Testing Programs. The University of Iowa.

Burstein, J., Kukich, K., Wolff, S., Lu, Ch., and Chodorow, M. (1998, August). *Enriching automated scoring Using discourse marking*. Paper presented at The Annual Meeting of the Association of Computational Linguistics. Montreal, Canada.

Charness, N. (1991). Expertise in chess: the balance between knowledge and search. En K. A. Ericsson and J. Smith, eds., *Toward a general theory of expertise*. N. Y.: Cambridge University Press, 39-63.

Chase, W. G. and Ericsson, K. A. (1981). Skilled memory. In J. R. Anderson, ed., *Cognitive skills and their acquisition* (pp. 141-189). Hillsdale, N.J: Erlbaum.

Chen, Ch.; Lee, S.Y.; Stevenson, H.W. (1996). Long-term prediction of academic achievement of American, Chinese, and Japanese adolescents. *Journal of Educational Psychology, 88*, 750-759.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*, 93-106.

Coleman, W., and Cureton, E. E. (1954). Intelligence and achievement: The jangle fallacy again. *Educational and Psychological Measurement, 14*, 347-351.

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

Edwards, A.J. and Kirby, M.E. (1964). Predictive efficiency of intelligence test scores: Intelligence quotients obtained in grade one and achievement test scores obtained in grade three. *Educational and Psychological Measurement, 24*, 941-946.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S. E. (1997). Multicomponent response models. En W. J. van der Linden, and Hambleton, R. K. *Handbook of Modern Item Response Theory*. New York: Springer Verlag, Pp. 305-322.

Ericsson, K. A. (1996).The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson, ed., *The road to excellence. The acquisition of expert performance in the arts and sciences, sports and games*, Mahwah, N. J.: Erlbaum, 1-50.

Ericsson, K. A., and Smith, J. (1991). Prospects and limits of the empirical study of expertise: an introduction. En K. A. Ericsson and J. Smith, eds., *Toward a general theory of expertise*. N. Y.: Cambridge University Press, 1-38.

Fischer, G. H., and Seliger, E. (1997). Multidimensional linear logistic models for change. En W. J. van der Linden, and Hambleton, R. K. *Handbook of Modern Item Response Theory*. New York: Springer Verlag, Pp. 323-346.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*, 285-307.

Frederiksen, N., Mislevy, R., and Bejar, I. (1993). Tests theory for a new generation of tests. N.J.: Erlbaum.

Glaser, R. (1994). Instructional technology and the measurement of learning outcomes: Some questions. *Educational Measurement: Issues and Practice*. (original, 1963), *13*, 6-8.

Glaser, R., Lesgold, A., and Lajoie, S. (1985). Toward a cognitive theory for the measurement of achievement. In J. Ronning, J. Glover, and J. Witt, eds., *The influence of cognitive psychology on testing and measurement*: The Buros-Nebraska symposium on measurement and testing (vol 3, pp 41-85). Hillsdale, N. J.: Erlbaum.

Glaser, R., and Silver, E. (1994). *Assessment, testing, and instruction: Retrospect and Prospect* (CSE Technical Report 379). University of Pittsburg: CRESST/Learning Research and Development Center.

Herman, J. (1999). *What constitutes a quality assessment*? CRESST web site: http://www.cse.ucla.edu.

Holland, J. G., and Skinner, B. F. (1961). *Analysis of Behavior*. N. Y.: McGraw- Hill.

Johnsen, S.K. and Ryser, G.R. (1997). The validity of portfolios in predicting performance in a gifted program. *Journal for the Education of the Gifted*, 253-267.

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato´s problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240

Levine, M. (1976). The academic achievement test. Its Historical context and social functions. *American Psychologist, 31*, 228-238.

Linn, R. L. (1998). Validating inferences from National Assessment of educational progress achievement-level reporting. *Applied Measurement in Education, 11*, 23-47.

Linn, R. L.. (1999, February). *Assessment and accountability systems*. Paper presented at The National Conference, American Association of School Administrators.

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

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

Lukhele, R.; Thissen, D. and Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*, 234-250.

Lytton, H. (2000). Toward a model of family-environmental and child-biological influences on development. *Developmental Review, 20*, 150-179.

Ma, X. & Kishor, N. (1997). Attitude toward self, social factors, and achievement in mathematics: A meta-analytic review. *Educational Psychology Review, 9*, 89-120.

Martínez, M. E., and Bennett, R. E. (1992). A review of automatically scorable constructed-response item types for large-scale assessment. *Applied measurement in Education, 5*, 151-169.

Masters, G. N., and Wright, B. D. (1997). The partial credit model. En W. J. van der Linden, and Hambleton, R. K. *Handbook of Modern Item Response Theory*. New York: Springer Verlag, Pp. 101-122.

Mislevy, R. (1993). Foundations of a new test theory. In N. Frederiksen, R.J. Mislevy & l.L. Bejar (Eds.), *Test theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J. (1996). Evidence and inference in educational measurement. *Psychometrika, 39*, 439-483.

Mislevy, R. J. and Verhelst, N. (1990). Modelling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215.

Muraki, E. (1997). A generalized partial credit model. En W. J. van der Linden, and Hambleton, R. K. *Handbook of Modern Item Response Theory*. New York: Springer Verlag, Pp. 153-164.

Muraki, E. and Bock, R. D. (1996). *PARSCALE. IRT based test scoring and items analysis for graded open-ended exercises and performance tasks (version 3).* Chicago, IL: Scientific Software International.

Neisser, U., Boodoo, G., Bouchard, Jr., T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., and Urbina, S. (1996). Intelligence: Knowns and Unknowns. *American Psychologist, 51*, 77-101.

Niemi, D. (1999, February). *Assessment models for aligning standards and classroom practice*. UCLA Graduate School of Education and Information Studies. Center for the Study of Evaluation. National Center for Research on Evaluation, Standards and Student Testing. Conference of The American Association of School Administrators.

Nowell, A. and Hedges, L.V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance and extreme scores. *Sex Roles, 39*, 21-43.

O´Neil, Jr., H., Jr. and Schacter, J. (1997*). Test specifications for problem solving assessment* (Technical report 463). Center for the Study of Evaluation, Standards, and Student Testing. Graduate School of Education and Information Studies, Los Angeles, Ca: University of California.

Pasnak, R.; Willson, Q.A. and Whitten, J. (1998). Mild retardation, academic achievement and Piagetian or psychometric tests of reasoning. *Journal of Developmental and Physical Disabilities, 10*, 23-33.

Pirolli, P., and Wilson, M. (1998). A theory of the measurement of knowledge content, access, and learning. *Psychological Review, 105*, 58-82.

Ramsay, J. O., and Stout, W. (1997). *The impact of modern Statistics on Psychometrics: Accomplishments and Opportunitites*. Unpublished Manuscript.

Revuelta, J., and Ponsoda, V. (1999). Generación automática de ítems. In J. Olea, V. Ponsoda, and G. Prieto (1999). *Tests informatizados. Fundamentos y aplicaciones*. Madrid: Pirámide. Pp. 227-248.

Ruiz-Primo, M. A. (1998, September). *Science achievement: What we have learned from two alternative assessments*. Paper presented at conference CRESST.

Samejima, F. (1997). Graded response model. En W. J. van der Linden, and Hambleton, R. K. *Handbook of Modern Item Response Theory*. New York: Springer Verlag, Pp. 85-100.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and self-making inb methematics. En D. Grouws, ed., *Handbook for research on Mathematics teaching and learning*. N. Y.: MacMillan.

Shapley, K. S., and Bush, M. J. (1999). Developing a valid and reliable portfolio assessment in the primary grades: Building on practical experience. *Applied Measurement in Education, 12*, 111-132.

Shilpi, N. (1995). Capturing the power of classroom assessment. *Focus, 28*, 1-22. Educational Testing Service.

Singley, M. K. And Bennett, R. E. (1998, November). *Item generation and beyond: Applications of schema theory to mathematics assessment*. Paper presented at The Conference "Generating items for cognitive tests: Theory and Practice. Princeton, N. J: Educational Testing Service.

Stecher, B. M., Barron, S., Kaganoff, T., and Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey of Kentucky Teachers of Mathematics and Writing* (Technical Report 482). National Center for Research on Evaluation Standards and Student Testing (CRESST). Ca: University of California, Los Angeles.

Strong, S., and Sexton, L. C. (1996). Performance assessment for state accountability: Proceed with caution. *Journal of Instructional Psychology, 23*, 68-74.

Strong, S., and Sexton, L. C. (1997). Kentucky performance assessment of mathematics: Do the numbers add up?. *Journal of Instructional Psychology, 24*, 202-206.

Stumpf, H. & Stanley, J.C. (1996). Gender-related differences on the College Board's Advanced Placement and Achievement Tests, 1982-1992. *Journal of Educational Psychology, 88*, 353-364.

Sugrue, B. (1996). *Patterns of performance across different types of items measuring knowledge of items measuring of Ohm´s Law*. Technical report 405. CRESST/University of California, Los Angeles.

Throne, F. M.; Kaspar, J. C. and Schulman, J. L. (1965). The Peabody Picture Vocabulary Test in comparison with other intelligence tests and an achievement test in a group of mentally retarded boys. *Educational and Psychological Measurement, 25*, 589 595.

U.S. Department of Education (1996). *Pursuing excellence*. NCES 97-198, Washington D. C.: U.S. Government Printing Office.

Van der Linden, W. J. (1986). The changing conception of measurement in education and psychology. *Applied Psychological Measurement, 10*, 325-332

van der Linden, W. J. and Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer Verlag.

Wainer, H., and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H. and Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*, 103-118.

Wainer, H. and Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research, 64*, 159-195.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*, 461-481.

William, W.M. and Ceci, S.J. (1997). Are Americans becoming more or less alike?. *American Psychologist, 52*, 1226-1235.

Willson, V. L. (1989). Cognitive and developmental effects on item performance in intelligence and achievement tests for young children. *Journal of Educational Measurement, 26*, 103-119.

(Recibido:12/01/00; Aceptado:30/10/00)