Psicothema 2004. Vol. 16, nº 4, pp. 696-701
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Theory, methods, and practices in testing for the 21st century

Ronald K. Hambleton
University of Massachusetts at Amherst (USA)

The changes that can be expected in the theory, methods, and practices of testing and assessment are considered in this paper. Item response theory will replace classical test theory because it provides features such as model parameter invariance that can improve the construction and analysis of tests. Computers should replace paper and pencil testing because they provide more flexibility in test scheduling, scores can be provided immediately following test administration, and the capability to assess higher level thinking skills is available. New item types, stimulated by the availability of computers for testing, can be expected to increase the validity of testing. Finally, for each major change, challenges remain to be overcome and some of these are described in the paper.

Teoría, métodos y práctica de los tests en el siglo XXI. En este trabajo se analizan los cambios que se pueden esperar en el futuro en cuanto a la teoría, los métodos, la práctica de los tests y la evaluación en general. La teoría de respuesta a los ítems reemplazará a la teoría clásica de los tests, pues aporta ventajas tales como la invarianza de los parámetros de los modelos que pueden mejorar la construcción y análisis de los tests. Los ordenadores van a desplazar a los tests de papel y lápiz, dado que proporcionan una mayor flexibilidad en el uso de los tests, las puntuaciones están disponibles inmediatamente tras la aplicación de las pruebas, y permiten la evaluación de aptitudes mentales superiores. Cabe esperar que los nuevos tipos de ítems surgidos por la disponibilidad de los ordenadores mejoren la validez de los tests. Finalmente, por cada uno de estos grandes cambios aparecen retos que hay que superar, algunos de los cuales se tratan en este artículo.

Trying to predict specific future directions in the field of educational and psychological testing is difficult because of the speed with which both testing methods and computer technology are advancing. When I think back to my days as a beginning graduate student in 1966 at the University of Toronto in Canada, multiple-choice items and essay questions dominated the testing field; classical test theory was firmly in place as a framework for test design and analysis; statistical methods were limited and there were no statistical packages such as SAS and SPSS, and no IRT software; and all technically oriented students moved around the university campus with computer cards in their pockets or computer boxes in their arms on their way to and from the university computer center. Data analysis too was slow and error prone with big clunky calculators. I checked several 1966 issues of the Journal of Educational Measurement in preparing this lecture and found no more statistical sophistication in the articles than what we expect today of students with two semesters of statistics in our graduate programs in psychology and education.

Today, emphasis in large-scale testing and credential exams is moving away from selected-response formats such as multiple-choice items, to extensive use of performance-based assessments—to aid in the assessment of problem solving, critical thinking, writing, etc.; modern test theory (i.e., item response theory) is now replacing classical test theory as a framework for test development and analysis; the focus in data analysis is on multivariate procedures; and statistical modeling of test data is often complex involving item response theory (Hambleton, Swaminathan and Rogers, 1991; Lord, 1980), generalizability theory (Brennan, 2001), and structural equation modeling (Byrne, 1998). And, as I was preparing my lecture for the Doctores Honoris Causa Ceremony at the University of Oviedo, I realized that I had access to more statistical power with my personal computer, than the University of Toronto had for 25,000 students in 1966 when I began my doctoral studies. Finally, I have that same computer power in my home, my office, on the airplane that I flew over on for the ceremony, and even in my car, if I choose to work there with my personal computer! And with my wireless telephone I can be hooked up to my personal computer, do email, and share professional materials with colleagues all over the world.

No one could have predicted these changes that have occurred in the last 30 years –(1) new test theories, (2) the transition of testing from paper and pencil to the computer, and (3) the introduction of many new item types for testing. Now, many of us complain while we wait for our computer to power up or if we have to wait a few seconds for our computers to respond. I thought I could not be more happy than I was with my computer, and then I discovered high speed internet connections. My enjoyment and productivity have increased considerably. We communicate with our colleagues on the internet—rarely do we send letters or make telephone calls

Fecha recepción: 11-2-04 • Fecha aceptación: 18-5-04
Correspondencia: Ronald K. Hambleton
School of Education
University of Massachusetts at Amherst
Amherst, MA 01003 (USA)
E-mail: rkh@educ.umass.edu

anymore. It is a different world today than in 1966 when I began my doctoral studies in psychometric methods, and computers are primarily responsible for the changes taking place.

Though I should be hesitant to make psychometric predictions of any kind based up my experiences since 1966 (I will predict that Real Madrid and Barcelona will remain as world soccer powers during my lifetime), a number of changes might safely be predicted in the next 20 years because they are already having an impact on the testing field. In this paper, focus will be centered on three areas where changes are likely to take place and impact considerably on educational and psychological testing, selection testing, and credentialing examination practices: (1) test theory, (1) computers and testing, and (3) new item formats and scoring models. I'm very pleased to note too that Spanish psychologists along with American, Dutch, and Australian psychologists have been among the most productive and influential with these new developments. The work of Muñiz (e.g., Muñiz, 1990, 1996), Navas (2001), Olea, Ponsoda and Prieto (1999), Ana Delgado, Alberto Maydeu, Pedro Prieto and many other psychometricians from Spain has been very important in the development of test theory and testing methods, and that work now is having an international impact. It's common today to pick up an international measurement and statistics journal and see a contribution from a Spanish psychometrician. This, too, is a change from when I began my graduate studies in 1966.

Transition from Classical to Modern Test Theory

Many psychologists have seen occasional references to the Rasch model, latent trait theory, item response theory, item characteristic curves, computer adaptive testing, etc. in popular psychological testing texts, test manuals, and journals. These new psychometric terms are associated with modern test theory, known as «item response theory.» And very soon, psychologists who do not know about item response theory will be in the minority, and very much at a disadvantage in their research. In Spain, today, there are books on general testing practices (Navas, 2001), on test theory (Muñiz, 1990, 1996), and computer-based testing (Olea, Ponsoda and Prieto, 1999). Interest in psychometric methods is growing rapidly due to the expanded uses of educational and psychological assessments and the corresponding need for test score validity.

Until recently, classical test theory has provided the statistical underpinnings for both educational and psychological tests (Gulliksen, 1950). While popular psychological testing books such as those prepared by Thorndike and Hagen, Anastasi, and Cronbach, in the United States do not provide the relevant test theory and derivations, all of the popular measurement formulas and approaches for constructing tests, evaluating tests, and interpreting scores that appear in these books (e.g., Spearman-Brown formula, standard error of measurement, corrections for score range restrictions) are derived from classical test theory.

Despite the usefulness of classical test theory and models in psychometric methods, shortcomings in the basic theory underlying psychological testing and measurement procedures for test construction have been recognized for over 50 years. Gulliksen (1950) wrote about some of these problems in his classic test theory text, and it is not surprising that one of his own students was among the first to advance the era of modern test theory (Lord, 1952).

One shortcoming of classical test theory is that classical item statistics –item difficulty and item discrimination– depend on the particular examinee samples from which they were obtained. That is, test items look easy when administered to bright examinees, and harder when administered to less capable examinees. A consequence of this dependence on a specific sample of examinees is that these item statistics are only useful when constructing tests for examinee populations that are similar to the sample of examinees used in calibrating the test items. Unfortunately, the vagaries of field-testing of items is such that one cannot always be sure that the population of examinees for whom a test is intended is similar to the sample of examinees used in obtaining the item statistics. Item statistics that are independent of the particular sample of examinees where they were obtained would be preferable. Item statistics that are *invariant* over examinee samples is one of the goals of modern test theory.

Not only are popular classical item statistics used in test development sample dependent, but so are important test statistics such as test reliability and validity. Test reliability is higher when estimated in heterogeneous samples of examinees rather than in more homogeneous samples of examinees. Correction factors are often used to adjust reliability estimates for this problem but the fact is that the dependence of reliability indices on the choice of examinee sample is troublesome. Again, test statistics independent of particular examinee samples would be desirable.

A second shortcoming of classical test theory is that comparisons of examinees on the test score scale are limited to situations where examinees are administered the same (or parallel) tests. The seriousness of this shortcoming is clear when it is recognized that examinees often take different forms of a test When several forms of a test that vary in difficulty are used, examinee scores across nonparallel forms are <u>not</u> comparable unless one makes use of equating procedures that are complex to implement in practice, especially with classical equating methods (Kolen and Brennan, 1995).

There are many situations where the use of non-equivalent tests are of interest. Out-of-level achievement testing in schools is one example (that is, for example, administering grade 4 level tests to students who are struggling with grade 5 content). More effective administration of aptitude tests, personality tests, and quality of life tests is another example. Testing time can be cut in half by adapting a test to the examinee (see, for example, Mills, Potenza, Fremer and Ward, 2002; van der Linden, & Glas, 2000; Wainer, et al., 2000). Starting examinees at different points in an intelligence test based on some prior information about each examinee is another example. But these examples create a problem, and that is examinees who have taken different forms of the test need to be compared to each other, or to a norm group who took a different version of the test. As test scores are sample dependent, that is test scores depend on the set of items administered, they are not an adequate basis for score reporting or using norms tables. How can two examinees be compared based on their test scores, when the tests themselves may differ substantially in difficulty? This is one of the fundamental problems that classical test theory cannot solve easily.

A computer-adaptive test, called a «CAT» is another excellent example of the problem of item dependent scores. A CAT is a test administered by a computer, where the items administered are dependent on the examinee's performance on previous items: examinees who perform well receive harder items to complete; examinees who perform poorly receive easier items to complete. But, again, the non-equivalence of test forms makes comparisons among examinees or comparisons of examinees' test scores to passing scores difficult without the use of complex equating methods.

What is needed, if the goal is to tailor or adapt the administration of tests to examinees, is an approach to ability estimation that is *not* test dependent. The influence of the particular items on the test administered to the examinee needs to be accounted for. Frederic Lord and Harold Gulliksen from the Educational Testing Service in Princeton, New Jersey, and many other psychometricians in the 1940s and 1950s were interested in producing a psychometric theory that assessed examinees in a way that did not depend directly on the *particular* items that were included in a test. The idea was that an examinee may score high on an easy test or lower on a hard test, but there was a more fundamental ability that the examinee brings to any given testing situation that does not change as a function of the sample of items administered. It is that more fundamental characteristic of the examinee that is usually of interest to the psychologist and it is that more fundamental characteristic, referred to as a «latent variable», that is of interest in modern test theory. This construct of interest, whatever the ability score measures, is more fundamental than test score because ability scores, unlike test scores, do not change with the particular choice of items in a test. Still, they could change over time because of instruction, life changes, experiences, etc., and that would be acceptable, and even expected.

The purpose of item response theory (IRT) is to overcome the shortcomings of classical test theory by providing a reporting scale on which examinee ability (the construct measured by the test) is independent of the particular choice of test items that are administered. What began in the 1940s and 1950s as a goal of psychometricians, became reality 30 years later (see, Lord, 1980). By the early 1970s, the theory was developing nicely, computer software was available, and applications of IRT were beginning to appear. As we begin the 21st century, IRT is being used by test publishers, large testing agencies, test developers, agencies conducting the international comparative studies of educational achievement, and psychologists around the world to address technical problems such as the automated design of tests, the study of item bias, equating test scores, computer-adaptive testing, and score reporting (Hambleton and Pitoniak, 2002; Hambleton, Swaminathan and Rogers, 1991).

IRT, in its basic form, postulates that (1) underlying examinee performance on an test is a single ability or trait, and (2) the relationship between the probability that an examinee will provide a correct answer (or agree to a statement, in the case of a personality or attitude survey) and the examinee's ability can be described by a monotonically increasing curve or function. We would expect examinees with more ability to have a higher probability of providing a correct answer than those with less ability so this feature is highly desirable. Or, in the case of (say) an instrument measuring attitudes towards a topic, we would expect those persons with very positive attitudes to agree with a statement more frequently than those persons with less positive attitudes.

There is not time in this lecture to introduce the models, concepts, and assumptions of item response theory (see, for example, Hambleton, Swaminathan and Rogers, 1991; Lord, 1980). Suffice to say, within an IRT measurement system, ability estimates for an examinee obtained from tests that vary in difficulty will be the same, except for the expected measurement errors. Some samples of items are more useful for assessing ability, and therefore the corresponding errors associated with ability estimation will be smaller than item samples that are less optimal. But the ability parameter being estimated is the same across item sets unlike in classical test theory where the person parameter of interest, true score, is test dependent. Sample invariant ability estimates are of immense value in testing because tests can be matched to the ability level of examinees to minimize errors of measurement and maximize test appropriateness, while at the same time, comparisons in ability scores can be made because the ability estimates are **not** test dependent.

The concept that ability and item parameters do not change as a result of different samples of persons and items is known as *ability parameter invariance and item parameter invariance*, respectively. In theory, this is because when the item parameters are estimated, ability estimates are used in the item parameter estimation process (that is not the case in classical test theory). Also, when examinees' abilities are estimated, item parameter estimates are incorporated in that process (again, this is not the case in classical test theory). Both ability estimates and item statistics are reported on the same scale, so they look different from classical test scores and item statistics. Finally, IRT provides a direct way to estimate measurement error at each ability estimate (score level). In classical test theory, it is common to report a single estimate of error, known as the standard error of measurement, and apply that error to all examinees. Clearly, such an approach is less satisfactory than producing an error estimate at each ability score level.

IRT models such as the one-, two-, and three-parameter logistic models provide estimates of both invariant item and ability parameters. Both features are of considerable value to test developers because they open up new directions for assessment such as adaptively administered tests and item banking. Of course, the feature of *invariance* will not always be present. Item and ability parameter invariance will be obtained when there is (at least) a reasonable fit between the chosen IRT model and the test data. Not surprisingly then, considerable importance is attached to determining the fit of an IRT model to the test data (see, for example, Hambleton, Swaminathan and Rogers, 1991).

There are IRT models today to handle nominal, ordinal and equal-interval educational and psychological data: One-, two-, and three-parameter normal ogive and logistic models; partial credit and graded response models; cognitive component models; rating scale model; nominal response model, and many more. Multidimensional normal ogive and logistic models are available too. There are at least 100 IRT models in the measurement literature, and about 10 of these are receiving wide use today (see, Hambleton and Pitoniak, 1997, van der Linden & Hambleton, 1997).

*Challenges***.** The various applications have been sufficiently successful that researchers in the IRT field have shifted their attention from a consideration of IRT model advantages and disadvantages in relation to classical test theory to consideration of such IRT technical problems as goodness-of-fit investigations, model selection, parameter estimation with small samples, and steps for carrying out particular applications (e.g., automating the item selection process in test development). Certainly issues and technical problems remain to be solved in the IRT field, but it would seem that IRT technology is more than adequate at this time to serve a variety of uses in the testing field.

Paper and Pencil Testing to Computer-Based Testing

The biggest change in the next 20 years will be the administration of more tests at a computer (see, for example, Luecht, 1998; Luecht and Clauser, 2002; Wainer, 2000; van der Linden and Glas, 2000) . Actually, this is a safe prediction because the movement to computer-based testing (CBT) has been steady in recent years. Testing agencies are moving from simply producing a fixed form or parallel-forms of a test to various test designs at a computer. Testing agencies have either moved to some form of linear test (parallel-forms, or «linear on the fly» tests with each examinee receiving a unique set of items subject to content and statistical specifications) or they have moved to computer adaptive testing (CAT). Both extremes of computer based tests, in principle, allow for flexibly scheduling of tests for examinees and immediate score reporting, attractive features for examinees. It is in this area of psychometric advance that Spanish psychometricans have been especially productive and influential (see, Olea, Ponsoda and Prieto, 1999). CBTs also open up the possibility for the use of a number of new item formats for assessing higher level thinking skills (Irvine and Kyllonen, 2002; Zenisky and Sireci, 2002).

I am keenly interested in the implementation of new test computer-based test designs and much of my research time has been spent on this problem in recent years (see, for example, Hambleton and Xing, in press; Hambleton, Jodoin and Zenisky, in press). My favorite design at the present time is the multi-stage test (MST) design. Instead of individualizing the test by optimal selection of *each* test item as is done in CAT, in MST, optimal selection involves selecting a block of items called a «testlet» (perhaps 15 to 20 test items). This is a very useful design because it allows for individualizing or adapting the test to examinee ability, while at the same time allowing examinees to omit questions within a testlet and change answers until the time the examinee decides to move to the next testlet. These two features–omitting items, and changing answers- may not seem so important, but they are the major criticisms of examinees taking computer-adaptive tests such as the Graduate Record Exam (GRE) used in admission to graduate schools in the United States. The GRE-CAT was one of the first large scale examinations in the United States to move to the computer. That test today is administered all over the world to persons desiring to attend graduate schools in the United States.

Test committees have sometimes shown a preference for MSTs because they like being able to package items into testlets and checking them before they are used. With CAT designs, there are as many «tests» as examinees and test committees have much less control over the actual combinations of items that appear together on their tests. This makes some committees feel uneasy. At the same time, as items get more fully classified, and as software becomes more flexible, this advantage of MST over CAT is likely to disappear.

There are other test designs too that are being developed and studied. In one, «computerized mastery testing,» randomly parallel forms are constructed with their information functions centered at the passing score, and testing continues until an examinee can be confidently placed into a passing or failing category.

Test administration at a computer, often involves test development via computer software. Software is needed that can mimic test design committees and can handle complex content and statistical specifications. There are a number of software packages that can now handle test development based on principles from operations research and linear programming–see, for example, the research by van der Linden, Luecht, Stocking, Jones and others. But this software is based on statistical models and estimates of item parameters that contain error so they cannot be completed depended on to do what practical test developers might do. Our impression is that the software, generally, is easy to run, and produces good results. Even more general software that is user-friendly can be expected soon.

*Challenges***.** Research to develop new computer-based test designs that can shorten testing time, incorporate new item formats, maintain or improve decision consistency and decision accuracy, and be psychologically satisfying to examinees, is very much needed. More research is needed in modeling various content and statistical constraints, bank sizes, model misfit issues, item exposure controls, detection of over-exposed items, along with test design improvements, to see what can be learned for more effective implementation of tests or assessments at a computer. Some of these breakthroughs will undoubtedly come from Spanish psychometricians.

Multiple-Choice and Essay Item Types to New Computer-Based Item Types

It is in the area of new computer-based item types that we are going to see the most changes in the coming years. More than 50 new item formats have been counted (Hambleton and Pitoniak, 2002; Zenisky and Sireci, 2002) with many more variations on the way. These new formats involve everything from changing the materials presented to examinees, to the way examinees respond, to even the way examinees interact with the material. A couple of exemplary initiatives include the pioneering work of Randy Bennett at ETS along with many of his colleagues on new item types, the outstanding work of Brian Clauser, Ron Nungester, and many of their colleagues at the National Board of Medical Examiners with sequential problem solving tests in medicine, the pioneering and award winning work of Isaac Bejar, Henry Braun, and their ETS colleagues with the national examination to credential architects, and the work of Craig Mills, Gerry Melican, and Krista Breithaupt and their colleagues at the AICPA to build and to score complex simulation tasks for the national examinations for accountants. All of these innovations are expensive and time consuming to develop and implement but the research of these groups will lead to improvements in test development and scoring, and in time, costs will come down, and test validity will be increased.

There are less labor-intensive and inexpensive initiatives that appear promising for use in computer-based testing, too. None of these ideas is especially innovative and some have been in the testing literature for years, but the computer in one way or another enhances their use:

1. Multiple correct answers. We might call this «multiple true-false.» This format seems particularly attractive for tests where there may be multiple correct answers. It is easy enough to implement on a computer—it is much more difficult to implement with the standard answer sheets used with large administration paper-and-pencil tests.
2. Short answer. This format has a long history, but in the near future, short answers will be scored by computers (see, for example, new software coming from ETS called «e-writer»). This format removes the difficulty for item writers of producing four or five answer choices, and enhances the fidelity of tests by using an open-response format. ETS and other testing agencies in the US are routinely scoring essays today (see, Zenisky and Sireci, 2002).
3. Extended answer (essays). Of course this is an old format too but now that scoring essays can be done via computer—several testing agencies are adding extended answers and essays to their assessments—for example, the *Scholastic*

Assessment Test (SAT) and the *Graduate Management Admissions Test* (see Zenisky and Sireci, 2002).

4. Highlighting text. In the context of educational assessment, examinees might read a passage and then they could be asked to «highlight the sentence that conveys the main idea in the text.» One could think of thousands of variations. Again, this format, introduced to testing by Randy Bennett from ETS, reduces the necessity of creating answer choices, and increases the fidelity of the CBT.
5. Ranking. With this format, examinees might be asked to rank a set of options to a problem. There are times when this format may be preferable to selecting one choice, or choosing all of the correct choices. Sometimes this ranking notion is attempted with the multiple-choice format but often clues can be gleaned from the available choices to reduce the choices to one or two. It normally doesn't work very well with tests in a paper-and-pencil format. The format can have high fidelity for examinees.
6. Numerical responses. This is a good format with numerical problems administered via a CBT. It is a simple variation of number 2 above. No prompts are given via answer choices. It is easy to score and the troublesome, time-consuming task of creating plausible distractors can be by-passed.
7. «Drag and drop format.» As one example of this format in a credentialing context, one could imagine a medical candidate being asked to sort a list of medical diseases, and doing this by dragging these diseases from a list and dropping or placing these diseases into, say, three requested categories. This format opens up a number of interesting possibilities.

Over 50 formats have been identified in the measurement literature (Zenisky and Sireci, 2002). They may involve complex item stems, sorting tasks, interactive graphics, the use of both audio and visual stimuli, job aids (such as access to dictionaries), joy sticks, touch screens, sequential problem-solving, pattern scoring, and more. Clearly, many new item formats can be expected in the coming years. You only need to look to exams in the Information Technology industry (e.g., Novell and Microsoft) to see the possibilities. Readers are referred to more information on this topic in van der Linden and Glas (2000) and Irvine and Kyllonen (2002).

I have five cautions for testing agencies who want to consider new item formats in their tests: (1) be clear on the constructs that you want to measure (will these new item formats permit a better assessment of the content domains of interest? Where is the evidence?), (2) be concerned about the issue of fairness (Is there evidence that these new formats will not place members of international groups, minority groups, handicapped groups, etc. at a disadvantage?), (3) compile evidence of reliability and validity to support any testing changes (Is there technical evidence to show that these new formats do not reduce either reliability or validity of resulting placements of examinees?), (4) address practical considerations (Are the gains in test validity worth the extra costs and complexities of test delivery?), and (5) consider the possibility of coachability (Are any test changes going to be coachable, and hence influence test validity negatively?). All of my cautions are about test validity, and I would not endorse the use of any new item formats without research evidence to support them. At the same time, I remain very optimistic about the potential advantages of new item formats.

Given the cost of producing test items, and the difficulty that testing agencies have in building up their item banks, I am surprised that some good ideas for expanding item banks are not being tried or implemented—for example, item cloning and item algorithms (see, for example, Pitoniak, 2002). I expect more will be done in the coming years because computer based testing of examinees on a flexible time schedule requires expanded item banks or test score validity is likely to be adversely affected.

None of my thoughts are especially new or innovative, but I believe most or all of the ideas below will be adopted as the need for larger item banks is recognized:

1. Superficial changes to disguise items. (Even this minor change will be helpful in areas like standardized patient exams (SP) in medical credentialing where examinees may share information about the problems they encounter. For example, problems become known as «Bill with the heart problem» or the «Neurotic Betty» problem. Superficial name changes of the characters can essentially stop the value of sharing solution information from one examinee to another.)
2. «Milking» of good test items. (Here, good test items are spotted from their item statistics and changes are made to create families of test items. Sometimes the areas of an item where changes might be made are referred to as «item facets.») Based on my experiences in training test item writers, I know that this approach can work, and can be used to dramatically increase the size of item banks with quality items.
3. Development of algorithms for generating items. (Item writers start with a blank sheet of paper and try to sketch out a problem with lots of scope. This general approach will pay dividends to testing agencies that need to substantially expand the size of their item banks to maintain test security and test score validity.)

Challenges. I have seen important research using item clones, item generation rules, etc. to expand item banks (Pitoniak, 2002). How far ranging can these approaches be applied? How cost effective are they? Also, with more test items, there is an expanded need for field testing, and with field testing comes item exposure. One wonders about the possibilities of training item writers to estimate item statistics (to be used as priors with Bayesian estimation in IRT models) to reduce examinee field test sample sizes while maintaining valid item statistics.

Conclusions

The psychological testing field is expanding daily, and computer technology, and to a much lesser extent, cognitive science, are driving most of the advances. In preparing this paper I mentioned many of the changes since 1966, and mainly I wanted to address the next 20 years of advances. But you could say that the future is now! New computer-based test designs, an expanded number of item formats, more sophisticated scoring models (e.g., pattern scoring and testlet), and improved modeling of examinee performance, today, are offering the potential for more valid assessments. All of these changes should permit test fidelity to be higher, and permit the assessment of many skills in more appropriate ways than were offered by the multiple-choice item format. But costs will go up, and change simply because an innovation has face validity or «sizzle,» cannot be recommended or defended. Changes in the methodology of testing practices should follow from careful, systematic research–and I believe more research is needed to insure that innovations are valid prior to changes being made. We have sufficient controversy in testing today, we don't need more controversy because of ill-conceived innovations.

So I end with a very positive impression of the changes that are taking place in the theory, methods, and practices of testing, and believe that with a strong research base, the next generation of tests is going to be even better at meeting the needs of users.

Note

Doctores Honoris Causa Lecture at the University of Oviedo, Spain, January 17, 2003.

References

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Byrne, B.M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gulliksen, H. (1950). *Theory of mental tests***.** New York: Wiley.

Hambleton, R.K., Jodoin, M. and Zenisky, A. (in press). Computer-based test designs and item formats for the next generation of assessments**.** *International Journal of Testing*.

Hambleton, R.K. and Pitoniak, M.J. (2002). Testing and measurement: Advances in item response theory and selected testing practices. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology-volume 4* (3rd ed., pp. 517-561). New York: John Wiley and Sons.

Hambleton, R.K., Swaminathan, H. and Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hambleton, R.K. and Xing, D. (in press). Computer-based test designs with optimal and non-optimal tests for making pass-fail decisions. *Applied Measurement in Education*.

Irvine, S.H. and Kyllonen, P.C. (Eds.). (2002). *Item generation for test development*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kolen, M.J. and Brennan, R.L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph No. 7*. Psychometric Society.

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

Luecht, R.M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22*, 224-236.

Luecht, R.M. and Clauser, B.E. (2002). Test models for complex computer-based testing. In C.N. Mills, M.T. Potenza, J.J. Fremer and W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments***.** Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Muñiz, J. (1990). *Teoría de respuesta a los ítems*. Madrid: Pirámide.

Muñiz, J. (Ed.) (1996). *Psicometría*. Madrid: Universitas.

Navas, M.J. (Ed.) (2001). *Métodos, diseños y técnicas de investigación psicológica*. Madrid: Universidad Nacional de Educación a Distancia.

Olea, J., Ponsoda, V. and Prieto, G. (Eds.) (1999). *Tests informatizados: fundamentos y aplicaciones*. Madrid: Pirámide.

Pitoniak, M. (2002). *Automatic item generation methodology in theory and practice* (Center for Educational Assessment Research Report No. 444). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

van der Linden, W.J. and Glas, C.A.W. (Eds.) (2000). *Computer adaptive testing: Theory and practice*. Boston, MA: Kluwer Academic Publishers.

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

Wainer, H. et al. (Eds.) (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zenisky, A.L. and Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337-362.