European Journal of Psychological Assessment, Vol. 13, Issue 2, pp. 140–152 © 1997 Hogrefe & Huber Publishers

Use and Consequences of Assessments in the USA: Professional, Ethical and Legal Issues

Wayne J. Camara

The College Board, Princeton Junction NJ, USA

Tests and assessments in the USA have taken on additional burdens as their uses have been greatly expanded by educators, employers, and policy makers. Increased demands are often placed on the same assessment, by different constituencies, to serve varied purposes (e. g., instructional reform, student accountability, quality of teaching and instruction). Such trends have raised considerable concerns about the appropriate use of tests and test data and testing is under increased scrutiny in education, employment and health care. This paper distinguishes among the legal, ethical and professional issues recently emerging from the increased demands of assessments and also identifies unique issues emanating from computer-based modes of test delivery and interpretation. Efforts to improve assessment practices by professional associations and emerging issues concerning the proper use of assessments in education are reviewed. Finally, a methodology for identifying consequences associated with test use and a taxonomy for evaluating the multidimensional consequential outcomes of test use within a setting are proposed.

Keywords: ethics, legal issues, assessment, USA

Individuals are first exposed to tests and assessments very early in their school years in the United States. By the age of 18, assessments have already played a significant role in the life decisions of many young adults such as graduation, promotion and retention, college admissions, placement, and scholarship awards. Tests and assessments are also widely used in the psychological and educational screening of children and adults, for career and vocational assessment, for certification and licensing of individuals for a number of occupations, and for the selection and placement of workers within government and private sector organizations. Given the diverse and important use of tests and assessments, measurement professionals have become increasingly concerned with questions of validity, fairness, intended use(s) and consequences related to the appropriate use of educational and psychological assessments.

The ethical and legal conduct of lawmakers, celebrities, athletes and professionals from all areas (e. g., business, investment, marketing, law) have attracted headlines and the attention of mass media. In the U. S. and many European countries there has been a fixation on such ethical and legal issues involving the responsible conduct and obligations to the public in recent years. Attention has also focused on the use of tests and test results in education, employment, and health care settings (Berliner & Biddle, 1995). The real and perceived misuses of assessments and assessment results have become one of the most challenging dilemmas facing measurement professionals and test users today. Abuses have been widely reported in preparing students to take tests and in the use and misuse of data resulting from large-scale testing programs. Elaborate and costly test cheating operations have been disclosed by federal agencies (Educational Testing Service, 1996), test preparation services have employed confederates to allegedly steal large pools of items from computer-based admissions testing programs, instances of students and employees being provided with actual test items before an administration have been reported, as have unauthorized extension of time limits, falsification of answer sheets and score reports, and violations of the confidentiality of test data (Schmeiser, 1992). Misuses of test data in high stakes programs abound and the accuracy and marketing tactics of test publishers have been criticized in some (Sackett, Burris, & Callahan, 1989; Sackett & Harris, 1984).

Professional conduct and responsibilities in use of assessments can be ordered within three levels: (1) legal issues, (2) ethical issues, and (3) professional issues. The practices and behaviors within these three levels are certainly interrelated, yet this categorization is useful to initiate a discussion of concerns, the severity of inappropriate practices and behaviors and examples of assessment practices which are most likely to be questioned by professionals and the public.

This paper, focusing on the U. S., will first discuss the assessment practices and behaviors which raise professional, ethical and legal concerns. Second, the paper discusses efforts in addressing these concerns and the diversity among individuals using tests and test results. Third, a paradigm is proposed for identifying and evaluating consequences associated with test use. Unique professional issues emerging from the increased reliance on technology and computer-based testing are briefly reviewed. Finally, the variety of such issues directly relating to educational assessments are illustrated to provide a context for discussions of the technical, legal and professional issues involved in assessment.

Legal, Ethical and Professional Issues in Testing and Assessment

It is difficult to define the boundaries of and distinctions between professional, ethical and legal issues or concerns surrounding the development and use of tests and assessments. Legal, ethical, and professional issues form a continuum of standards for professional conduct in assessment and other areas. Laws and government regulations are legal mandates that affect all individuals living in a society. Ethical codes may range from enforceable to exemplary to educational principals that guide the professional behavior and conduct of members of any profession. Professional guidelines, principals and standards are also developed to educate and guide professionals in more technical activities. All three layers of regulations or standards exist in testing and assessment.

Laws and legal documents about testing and assessment are generally vague and ambiguous, but it is clear that where they exist, they have greatly influenced both professional standards of conduct and professional practices in assessment and testing. Government involvement and regulation of testing is most evident in personnel testing. But even here, laws and legal challenges to testing are limited to very specific domains. They address some issues (e. g., discrimination) and applications (e. g., employment testing) of assessment which have received widespread attention, while leaving many more common issues and concerns of test use unaddressed (e. g., validity of the measure when disparate impact is not present). Numerous federal and state laws and executive orders have implications on employment testing primarily through prescribed standards for equal employment opportunity (Camara, 1996), but also for the assessment of individuals with disabilities, the handling and retention of personnel records, and restrictions of the use of certain pre-employment techniques (e. g., Employee Polygraph Protection Act of 1988). The general consensus among industrial psychologists is that Civil Rights laws, which emanated in the 1960s, have been a major stimulus for improved pre-employment assessment practices. Employers became better educated about the technical, professional, and legal issues involved in the use of testing out of necessity, and while there is some evidence that regulations initially decreased use of employment testing, today they are used by a higher proportion of organizations than ever (Deutsch, 1988).

The first formal ethics code for any profession using assessments was adopted by the American Psychological Association (APA) in 1952. Eighteen of the more than 100 ethical principals from this Code (APA, 1953) addressed the use of psychological tests and diagnostic aids, and addressed the following issues of test use: (1) qualifications of test users (3 principles); (2) responsibilities of the psychologist sponsoring test use (4 principles); (3) responsibilities and qualifications of test publisher' representatives (3 principles); (4) readiness of a test for release (1 principle); (5) the description of tests in manuals and publications (5 principles); and (6) security of testing materials (2 principles).

Codes from the Canadian and British Psychological Associations came later, as did those from other European nations (Lindsay, 1996). In the past decade, many other professional associations have adopted ethical standards and professional codes which cover measurement and assessment issues. These trends have resulted from the increased public awareness of ethical issues, the variety of new proposed and actual uses for assessments, the increased visibility given to assessments for accountability purposes, and a commitment from the professions to safeguard the public (Eyde & Quaintance, 1988; Schmeiser, 1992). Ethical standards of the American Counseling Association, and APA are unique in that these associations maintain formal enforcement mechanisms that can result in member suspension and expulsion, respectively. In 1992, the American Educational Research Association (AERA) adopted ethical standards, followed in

1995 by the National Council of Measurement in Education's (NCME) Code of Professional Responsibilities in Educational Measurement. Several other organizations such as the Society for Industrial and Organizational Psychology (SIOP) and regional I-O organizations formally adopted APA's most recent ethical code for their members for educational purposes without any enforcement mechanisms.

Laws which affect testing, primarily strive to protect certain segments of the public from specific abuses. Ethical standards and codes attempt to establish a higher normative standard for a broader range professional behaviors. For example, APA's ethical standards note that:

> ... in making decisions regarding their professional behavior, psychologists must consider this Ethics Code, in addition to applicable laws and psychology board regulations. If this Ethics Code establishes a higher standard of conduct than in required by law, psychologists must meet the higher ethical standard. If the Ethics Code standard appears to conflict with the requirements of law, then psychologists make known their commitment to the Ethics Code and take steps to resolve the conflict in a responsible manner (APA, 1992, p. 1598).

Coinciding with this increased attention to ethical codes has been a dramatic increase in professional and technical standards for assessment issued which is described later. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985) are the most widely cited document addressing technical, policy, and operational standards for all forms of assessments that are professionally developed and used in a variety of settings. Four separate editions of these standards have been developed by these associations and a fifth edition is currently under development. However, numerous other sets of standards have been developed to address more specific applications of tests or aimed at specific groups of test users. Standards have been developed for: (1) specific uses such as the validation and use of pre-employment selection procedures (Society for Industrial and Organizational Psychology, 1987), integrity tests (ATP, 1990), licensing and certification exams (Council on Licensure, Enforcement, and Regulation, 1993; Council on Licensure, Enforcement and Regulation & National Organization for Competency Assurance, 1993), educational testing (Joint Committee on Testing Practices, 1988); (2) for specific groups or users such as classroom teachers (AFT, NCME, NEA, 1990), and test takers (Joint Committee on Testing Practices, 1996), and (3) for specific applications such as performance assessments, adapting and translating tests (Hambleton, 1994), and admissions testing (College Entrance Examination Board, 1988; National Association of Collegiate Admissions Counselors, 1995).

Professional standards, principals, and guidelines are more specific and generally oriented toward more technical issues to guide test users with specific applications and use of assessments. First, technical issues concerning the development, validation, and use of assessments are addressed in standards. Validity is the overarching technical requirement for assessments, however, additional professional and social criteria have been considered in evaluating assessments, such as: (1) how useful the test is overall, (2) how fair the test is overall, and (3) how well the test meets practical constraints (Cole & Willingham, 1997). These criteria are directed at both the intended and unintended uses and consequences of assessments. Existing standards guide test users in the development and use of tests and assessments; however, these standards may rarely reach and influence test users not associated with a profession.* For example, most employers are unaware of *the Principles for the validation and use of personnel selection procedures* (Society for Industrial and Organizational Psychology, 1987) and certainly the vast majority of educational administrators and policy makers who determine how to use tests and cite test results in making inferences about the quality of education have never viewed a copy of any of the various standards in educational measurement and testing.

Professional standards developed by groups such as APA and AERA do not appear in publications commonly read by employers, educators and policy makers. Many standards are written at a level where they may be incomprehensible to such individuals even if they had access to them. Finally, in many instances, members of the professional associations which develop and release standards themselves may not have and use copies of the standards and may have had little exposure to the standards and other new topics in testing, measurement, and sta-

* Often professional standards have been cited by courts in case law and have influenced assessment practices in these ways.

tistics through graduate training courses (Aiken, West, Sechrest, Reno, Roediger, Scarr, Kazdin, & Sherman, 1990). For example, the *Standards for Educational and Psychological Testing*, referred to as *"the Standards"* (AERA, APA, & NCME, 1985), which are the most widely cited professional standards for any form of testing and assessment, have total sales of 56,000 through 1996, while there are more than 120,000 members of APA alone (Brown, 1997).

Efforts to Improve Proper Use of Tests and Assessments

In the 1950s, when the first technical standards for testing were adopted in the United States (APA 1955, AERA, & NCMUE, 1954) the test user was considered to be a trained professional who conducted testing or disseminated results and interpretations to an individual. This classic definition of test user includes psychologists, physicians, counselors, personnel managers, and state or local assessment directories who generally have both some training and explicit job responsibilities for assessment. These test users seek to qualify for purchasing test materials, provide detailed interpretation of test scores, or represent a client organization in procuring and using assessments. The current version of the *Standards* (AERA, APA, & NCME, 1985) follows this line in defining the test user as someone who "requires the test for some decision making purpose" (p. 1). These individuals may best be termed "primary test users" because of their role and responsibilities. Today the concept of test user is much broader and often includes many different individuals with little or no training in measurement and assessment. However, there are secondary test users, especially in education, including policy makers, teachers, parents and the media, who often have no general training in assessment and no prescribed responsibilities for assessment. Some of these individuals can greatly influence and distort the general interpretation of assessment results, misuse assessments and results, and may have political incentives to selectively use results to support certain values or beliefs (Berliner & Biddle, 1995). Many of the most striking examples of test misuse are generated and supported by such secondary users. The more removed the test user is from the examine, the less familiar they are with the personal characteristics, academic abilities, or workplace skills of the individual tested, the more likely instances of test misuse will occur.

With the expanded uses and increased focus on assessment has come renewed criticism of the misuses and negative consequences of assessments. Professionals in measurement and testing are increasingly struggling with how to best improve proper test use and to both inform and influence an increasingly diverse group of tests users today who may have no formal training in testing and measurement uses, but still have legitimate claims for using test results in a wide variety of ways. Such groups have attempted to address the legal, ethical, and professional concerns with additional codes of conduct, technical standards, workshops, and case studies. However, most of these efforts rarely reach beyond members of the specific professional association or clients/users of a specific assessment product. Clearly such efforts are essential for improving the proper use of assessments and appropriate understanding of assessment results. Yet, these initiatives will not generally reach the secondary users who may insist on using assessment results as the sole determinant of high school graduation, rewards and sanctions to schools and teachers, and the primary indicator of equity, educational improvement or student achievement. Unfortunately, efforts which are aimed at only one segment of a much more expansive population of test users may not go far enough, fast enough for improving proper assessment practice.

Associations have attempted to cope with this new more expansive group of "secondary test users" by developing broader and simpler forms of standards, such as the Code of Fair Testing Practices in Education which basically condenses the primary standards from a 100 page document into a four page booklet which encourages duplication. Other efforts have been to work in collaboration with broader groups such as the National Education Association to develop codified guidelines or standards. However all such efforts have had no consistent impact across situations because there are few common linkages, different priorities and expectations for assessments, and little common understanding between primary and secondary test users. Relatively few efforts have been focused on undergraduate and graduate programs which train teachers and measurement specialists. Because universities and colleges differ in types of programs offered, the title of courses and course sequences, and even the departments which such programs are housed in, targeting and reaching educational programs broadly presents a number of substantial logistical obstacles. Often it is difficult to identify the faculty and administrators responsible for such programs and to effect systematic changes in their training programs.

For these and other reasons, Haney and Madaus (1991) state that test standards have had little direct impact on test publishers practice and even less impact on test use. They note that professional codes and standards primarily serve to enhance the prestige, professional status, and public relations image of the profession rather than narrow the gap between standards and actual practice. How do we resolve these issues? Given the increased social policy implications of testing, some have argued that greater legal regulation, litigation, or enforcement of technical standards by independent auditing agency present some potential mechanisms for reducing the misuse of assessment practices (Haney, 1996; Haney, Madaus, & Lyons, 1993; Madaus, 1992). However, such mechanisms may have little impact on many of the most visible misuses of assessments because it is often legislative and executive branches of state and federal government who advance expanded and often inappropriate use of assessments. Because test use is so expansive and abuses are so diverse, solutions which address only one element or one audience (e. g., test developer, teacher) may not be equipped to resolve the majority of instances where assessments are misused.

Consequences of Testing and Assessment

A more thorough understanding and consideration of potential consequences of test user can substantially reduce inappropriate uses and the resulting ethical issues. When consequences are discussed we are often reminded of the exclusively negative consequences resulting from test use:

- adverse impact on minorities and women
- discouraging and ‘institutionalizing’ failure for individuals
- teaching to the test and limiting curriculum and learning
- reinforcing tasks requiring simple rote memory at the expense of more complex cognitive processes required for success in the real world
- creating barriers
- tracking individuals into classrooms and jobs offering fewer challenges and opportunities

There are also often positive consequences when validated assessments are appropriately used:

- merit as a guide for decision making (selecting the most qualified candidate or making awards based on relevant performance)
- efficiency (relatively quick and effective means of collecting a large amount of data across a range of skills/competencies)
- quality control (certification or licensure)
- protection of the public (negligent hiring for critical occupations)
- objectivity for making comparisons and decisions among individuals or against established criteria
- cost effectiveness and utility

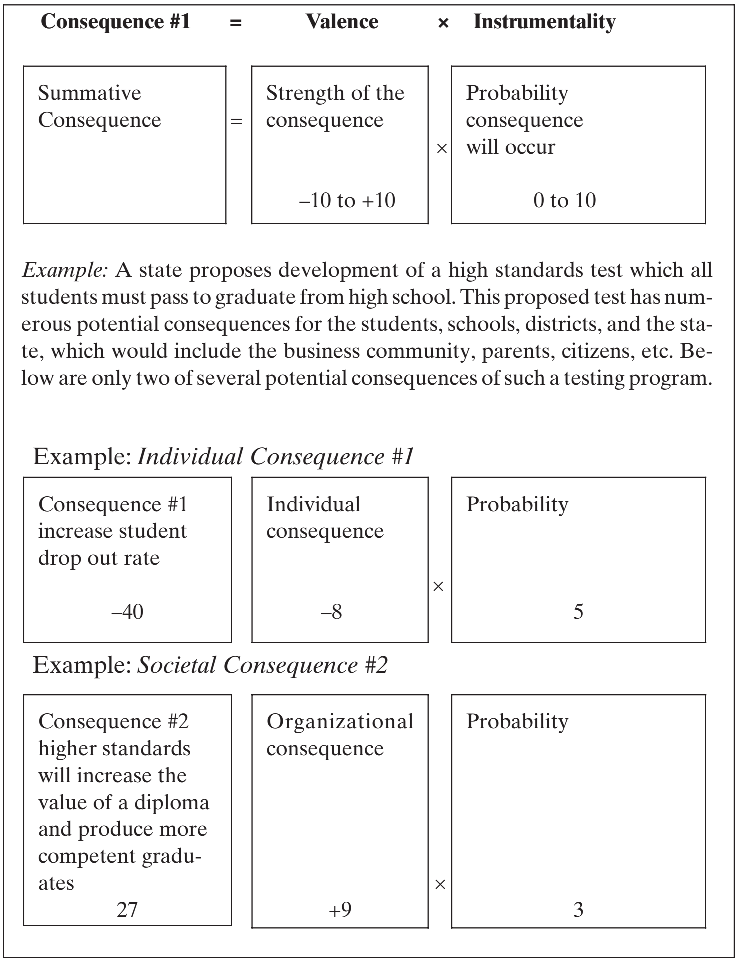
Consideration of how social ramifications of assessments affect validity has been summarized by Cronbach (1988) who stated that validity research is essentially a system which considers personal, institutional, and societal goals as they relate to inferences derived from test scores. If validity is established through evidence that supports inferences regarding specific uses of a test, then intended and unintended consequences of test interpretation and use should be considered in evaluating validity (Messick, 1989). Test developers and test users need to anticipate negative consequences that might result from test scores, potential corruption of tests, negative fallout from curriculum coverage, and how teachers and students spend their time (Linn, 1993). While there is some consensus that the consequences of test use must become an important criterion in evaluating tests within education, this view is not generally held in other settings (e. g., personnel, clinical).

Before consequences can be integrated as a component in evaluating assessments, a taxonomy or model is required. Such a taxonomy must consider both positive and negative impacts and consequences associated with test use. The impact, consequences, and feasibility of alternative procedures (e. g., biographical data, open admissions vs selection). Further complicating such a taxonomy is the knowledge that different stakeholders will have widely differing views on these issues. After the consequences have been identified, their probability of occurrence, the weight (positive or negative) associated with each consequence, and the level at which the consequence occurs (i. e., individuals, organizations, or society) must be determined.

This taxonomy borrows from terminology and processes from expectancy theory (Vroom, 1964) where the weight of the consequences are similar to the “valence” and the probability is related to the

Table 1. Paradigm for evaluating the consequential basis of assessments.

Consequence	Individual (e. g., student)	Organization (e. g., school)	Societal (e. g., community)
Positive			
Harmful			
Summative			


Consequence #1 = Valence × Instrumentality
Summative Consequence
=
Strength of the consequence
–10 to +10
×
Probability consequence will occur
0 to 10
Example: Individual Consequence #1
Consequence #1 increase student drop out rate
–40
Individual consequence
–8
×
Probability
5
Example: Societal Consequence #2
Consequence #2 higher standards will increase the value of a diploma and produce more competent graduates
27
Organizational consequence
+9
×
Probability
3

Example: A state proposes development of a high standards test which all students must pass to graduate from high school. This proposed test has numerous potential consequences for the students, schools, districts, and the state, which would include the business community, parents, citizens, etc. Below are only two of several potential consequences of such a testing program.

Figure 1. Computation of the summative consequence for each potential consequence associated with test use. Two examples of how specific outcomes of a high standards graduation test can produce summative consequences for arriving at a consensus regarding the potential positive and harmful consequences of assessment. For each of the possible consequences, the summative consequence would be summed by level (individual, organization, societal) to arrive at an overall index of the potential consequences of an assessment for the level. A consensus process would be employed to develop these values and to determine the overall desirability of the proposed assessment

"instrumentality." Proposed steps (adopted from, Camara, 1994) in determining the consequences of test use include:

1. Identify the intended consequences and objectives of assessment
2. Identify subject matter or content experts or develop an alternative consensus process
3. Identify potential intended and unintended consequences to individuals and organizations through a review of the literature, interviews or focus groups with key stakeholders
4. Determine the level that each consequence occurs (does it impact individuals, organizations, or society?)
5. Determine the probability of occurrence for each consequence (e. g., instrumentality)
6. Determine the strength or weight of each consequence (e. g., valence)
7. Employ a consensus process to determine the summative consequences of different aspects of test use on individuals, organizations and society.

Ideally the test developer, test users, and other key stakeholder groups would consider these issues before embarking on a new or revised testing program. Most consequences will have multiple impacts on individuals (e. g., test taker, teacher), organizations (e. g., schools, business), and society (e. g., community, state). Steps 5 and 6 require individuals, often with very diverse views, to arrive at a consensus or common judgment about the probabilities and strength (and direction) of consequences. The literature on standard setting may be of assistance in structuring a more explicit process.

Table 1 illustrates how consequences may be identified and classified through a consensus process. A list of potential consequences would be developed and classified within each of the nine boxes (step 4). Once all potential consequences are identified, each consequence is fully evaluated to determine its valence and instrumentality as illustrated in Figure 1. Step 7 in the process would have key stakeholders determine the overall summative consequences on individuals, organizations and society before a final decision is reached on the desirability and appropriateness of an assessment program or proposed use for assessments.

Such a taxonomy would not ensure that test misuse is minimized, but it would help to raise awareness of the diverse range of issues that emerge across different stakeholders and constituency groups who are involved in a high stakes assessment program. The absence of literature proposing models or taxonomies to identify and contrast consequences associated with test use leaves the test developer and user with little or no guidance in improving professional conduct and appropriate use of assessments.

Concerns Arising from Technology and Computed-Based Testing

Professional and ethical concerns about the appropriate use of testing have dramatically increased over the past few decades as the community of test users has increased and the use of assessments has expanded. Most recently technological innovations have given rise to a number of new and unique professional and ethical challenges in assessment.

Matarazzo (1986), and later, Eyde and Kowal (1987), identified many of the unique concerns created by the use of computerized clinical psychological test interpretation services. For example, Matarazzo noted that variation in clinical interpretation is often the rule rather than the exception, requiring much greater time and effort for clinical judgment and interpretation of computer-generated clinical interpretations than is usually the case. "Specifically, two or more identical soil readings, blood chemistries, meteorological conditions or MMPI profiles may require different interpretations depending on the natural or human context in which each is found ... use of the same 'objective' finding (e. g., an IQ of 120 or a '27' MMPI codetype) may be quite different if the 'unique' patient is a 23-year-old individual being treated for a first acute, frankly suicidal episode than if the 'unique' patient is a 52-year-old truck driver ... applying for total disability" (Matarazzo, 1986, pp. 20–21). Eyde and Kowal (1987) explained that computer-based testing provides greater access to tests and expressed concern about the qualifications of such expanded test users.

Technological innovations and the increased pressure for accountability in health care services may also be creating a different demand and market for clinical and counseling assessments. Assessments in all areas can be and are delivered directly to consumers. The availability of a take-home CD-ROM IQ test for children and adults, marketed to the general public or by Pro-Ed, a psychological testing and assessment publisher, has raised these same ethical and professional issues for psychologists. The CD-ROM test comes with an 80-page manual which informs parents of some of the theories of testing, how to administer the test, and how to deal with test results (*New York Times*, January 22, 1997). In such instances when tests are delivered by the vendor directly to the test taker there is no traditional test user. The test taker or their parents, who have no training and little knowledge of testing, must interpret the results, which increases the risk of test misuse.

Computer-adaptive testing (i. e., assessments in which the examine is presented with items or tasks matched to his or her ability or skill level) is increasingly used for credentialling and licensing examinations and admissions test programs today. Several unique concerns arise even when computer-based tests are administered under controlled conditions, such as in the above instances. First, issues of equity and access arise because these computer-based testing programs often charge substantially higher testing fees, which are required to offset the additional expenses incurred for test development and delivery, and often have more limited geographical testing locations. Second, familiarization with technology and completing assessments on computer may be related to test performance. Research has demonstrated that providing students with practice tests on disk in advance of testing, and tutorials at the beginning of the test administration are important in reducing anxiety and increasing familiarization with computer-based testing. Third, differences in the format, orientation, and test specifications of computer-based testing may affect the overall performance of individual test takers. Russell and Haney (1997) note that students who use computers regularly perform about one grade level worse if tested with a paper-and-pencil test than with a computer-based test. Students completing computer adaptive tests may also become more frustrated and anxious as they receive far fewer "easy items" since item selection algorithms are designed to match items with the level of each test taker's ability — resulting in more items that are perceived as "hard" by the test taker. Additionally, computer-based tests generally do not permit test takers to review items previously answered, as is common on paper-and-pencil tests. Additional rules are required for students who omit a large number of items and disclosure of test forms. Computer adaptive testing could be manipulated by test takers or coaching schools if some minimum threshold of item completion were not required. Because exposure of items is a major risk with these tests disclosure of test forms can not be as easily accommodated as with paper-and-pencil tests (Mills & Stocking, 1996). These and other distinctions associated with computer-based testing raise additional professional issues for test users, test developers, and test takers. As Everson (1997) notes, convergence of new theories of measurement with increased technology presents many opportunities for improved assessment frameworks, but also raises additional professional and ethical issues concerning assessment.

Fees for computer-based testing programs have generally been running between 300 to 600% higher than the fees for the same paper-based tests. Currently, test takers will receive immediate score reports and slightly more precise measurement accuracy at their level, but little additional benefits from the higher test fees. The few national programs offering computer-based testing programs have either eliminated (or plan to eliminate) the paper-and-pencil or raised fees on the paper-based test to ensure adequate volume for the higher priced computer-based product. Until additional advantages are realized from computer-based tests, business practices of replacing a lower priced test with one that is three to six times as costly for the test taker should be questioned.

Educational Assessment Today: Legal, Ethical and Professional Concerns

As tests are increasingly used for distinct and multiple purposes negative consequences and misuse are more likely to emerge. The use of performance assessments and portfolios in high stakes assessment programs can also raise additional issues about standardization and fairness. Nowhere are these concerns more evident than in educational assessment today.

In the past decade, there have been expanded expectations for assessments to not only measure educational achievement but to bring it about. Assessments are increasingly viewed as tools to document the need for reform by holding schools and students accountable for learning, and also as leverages of reform (Camara & Brown, 1995; Linn, 1993). President Clinton has proposed development of national assessments for all students in reading and mathematics by 1999 and called on all schools to measure achievement of their students in these and other areas. Currently, forty-eight of fifty states in the U. S. currently have in place or are developing large-scale educational assessments to measure the performance of their students. In some states these tests are used for high stakes purposes such as issuing a diploma or rewarding/sanctioning schools, districts, and even individual teachers. State and local boards of education and state and local departments of education translate test performance to make decisions about schools and/or individuals. School administrators come under increased pressure in such high-stakes testing programs to defend instructional practices and student achievement. Classroom teachers who administer the assessments and increasingly view them as a driving force for instructional change and educational reform also have a role in such assessment programs. Parents, students, and the general public who demand improved quality in education, business leaders who are often critical of graduates for lacking appropriate workplace skills, higher education which finds an increasing proportion of incoming students requiring remedial instruction, and policy makers who must respond to all these diverse stakeholder groups represent many different types of secondary test users.

Dissatisfaction with standardized assessments is also greatest in education because of their perceived negative consequences on learning and instruction. The performance assessment movement has strong support both within education and educational measurement and has not become another educational fad as some had predicted. Several large assessment programs had sought to replace their standardized testing programs with fully performance-based or portfolio systems. Today it appears that the "model" state assessment program will combine such constructed response tasks with more discrete, selected response (e. g., multiple choice, grid-in's) test items. Employing multiple measures allows educators to gain the benefits of more indepth and applied performance tasks that increase curricular validity, as well as increased reliability and domain coverage that selected response items offer. However, a number of legal, ethical and professional concerns emerge with any high stakes assessment program whether the decisions made primarily affect the student or the school.

Single assessments, either norm-reference multiple choice assessments or more performance-based assessments, do not well serve multiple, high-stakes needs (Cresst, 1995). Often key proponents of large-scale assessments support multiple uses, but actually have very different priorities given these uses. Kirst and Mazzeo (1996) explain that when such a state assessment system moved from a design concept to becoming an operational testing program it became clear that not all the proposed uses and priorities for the design could be accommodated. When priorities of key stakeholders could not be met, support for the program decreased.

Phillips (1996) identified legal criteria which apply to such expanded uses of assessments for high stakes purposes. These criteria have been modified and supplemented with several additional criteria which reflect a range of issues:

Adequate advance notification of the standards required of students. To ensure fairness, students and parents should be notified several years in advance of the type of standards they will be held to in the future. Students and teachers should be provided with the content standards (knowledge and skills required) and performance standards (level of performance). Sample tasks, model answers, and released items should be provided and clear criteria should be established when high stakes (e. g., graduation) uses are associated with the test.

Evidence that students had an opportunity to learn. The critical issue is whether students had adequate exposure to the knowledge and skills included on the assessment or whether they are being asked to demonstrate competency on content or in skills that they were not exposed to in school. Phillips (1996) notes that such curricular validity can often be demonstrated through survey responses from teachers that ensure students had on average more than one opportunity to learn each skill tested.

Evidence of opportunity for success. This challenge emerges when major variations from standardization occur. This assumes that all students are familiar with the types of tasks on the assessment, the mode of administration (e. g., computer-based testing), have the same standardized administrative, scoring procedures, and equipment (e. g., some students have access to a calculator or superior laboratory equipment in completing the assessment), and that outside assistance (e. g., group tasks, student work produced over time where parents and others could unduly offer assistance) could not affect performance on the assessment. Variations in these and other conditions can present an unfair advantage to some students.

Assessments reflect current instructional and curricular practices. If assessments are designed to reflect exemplary instructional or curricular practices, as is often the desire of educators who hope to use the assessment to drive changes, which are not reflected in the actual practices for many schools, a fundamental fairness requirement may not be met. The same challenges could be brought where teachers do not receive the professional development to implement new instructional or assessment practices (e. g., use of a graphing calculator) that are required on the assessment or in end-of-course assessments where the teacher lacks appropriate credentials for the subject area.

While these concerns apply to most educational assessments, they move from professional issues to legal and ethical concerns when assessments are used to make high stakes decisions. Additional ethical and professional issues which have been associated with various high stakes educational assessments may also affect other types of testing programs in other settings. Only a few of these issues are briefly addressed below.

Overreliance or exclusive reliance on test scores. Test performance should be supplemented with all relevant and available information to form a coherent profile of students when making individual high stakes decisions (e. g., admissions, scholarships). Student performance on tests should be interpreted within the larger context of other relevant indicators of their performance. In admissions decisions, students grades, courses, and test scores are generally all considered, with supporting information on personal qualities and other achievements. When testing has been repeated, performance on all administrations will permit individuals to identify any particular aberrations; less weight should generally be assigned to that particular test score or other indicator in these instances. Similar errors occur when individuals overinterpret small score differences between individuals, groups, or schools.

Cheating and "teaching to the test." There have been numerous examples of individuals cheating on high stakes tests. In addition, several instances where educators and other test users have been accused of systematic efforts of cheating (e. g., excessively high erasure rates on students papers, disclosure of answer keys to job incumbents on promotional exams) have received national attention, with some estimates that over 10% of test takers are cheating on high stakes tests (Fairtest, 1996). Because test scores are used as an indicator of school quality, school performance influences property values, school funding, and school choice — creating added incentives to increase school and district test scores by any means possible. These pressures often result in teaching to the test according to many educators and this common criticism of standardized tests. It is such negative consequences and the prospect for improved schooling that has caused the impetus for performance assessment, not the desire for better measurement for its own sake (Dunbar, Koretz, & Hoover, 1991).

Consideration of the cultural and social experiences of the test taker. Students bring their prior social, and cultural experiences with them when they participate in the class, compete in a sporting event, or complete an assessment. For many students the cumulative effect of these experiences may be to emphasize certain behaviors, skills or abilities that are less similar to those required of the assessment. The greater the similarity of an individual's socioeconomic and cultural background to that of the majority population, the better his or her test performance will generally be (Helms, 1992). Additional efforts are required to both ensure that all students are familiar with the types of tasks on the assessment and to ensure that divergent skills and abilities are considered in the construction and validation of assessment programs. Sensitivity to cultural, ethnic, gender, and language differences are required when interpreting results from assessments or other measures for individual students. Similarly, differences in these and other demographic variables must be considered when making simplistic comparisons among schools, districts, and other units. When these issues are not adequately considered by test developers and test users serious professional and ethical issues arise.

Exclusion of students from large-scale testing programs. Most large-scale national assessment programs which use aggregate level data (school, district, state) to monitor educational progress and permit comparisons systematically exclude large proportions of students with limited English proficiency and disabilities (McGrew, Thurlow, & Spiegel, 1993) Often school staff determine which students may be excluded from such national and state testing programs and there is variation across schools in the exclusion rates and application of criteria for excluding students. Paris, Lawton, Turner, and Roth, (1991) have also demonstrated that "low achievers" are often excluded by some schools or districts which would have the effect of artificially raising district test scores. Such practices introduce additional error into analyses, complicate accurate policy studies, affect the rankings resulting from the test data and introduce a basic unfairness in the use of test data (National Academy of Education, 1992).

Use of test scores for unintended purposes. Many of the most visible misuses of tests occur when scores are used for unintended purposes (Linn, Baker, & Dunbar, 1991). This occurs with state comparisons of unadjusted SAT or ACT scores, when results from state assessments are used as indicators of teacher competence, and when test results become the primary basis for inferences concerning the relative quality of learning or education among different schools or geographical regions. Test scores will always be considered an important indicator of learning. However, test users must become more aware of the extraordinary limitations and weaknesses of placing undue weight on test scores in such situations. Many state reports cards have attempted to provide public reports on the quality of education by examining a range of criteria (e. g., safety, learning, continuation in higher education, gainful employment, student honors) with a range of indicators that extend beyond test scores. As the test user becomes increasingly removed from personal knowledge of the examine, or less familiar with the units (e. g., schools, districts) of comparison, instances of mismeasurement and test misuse will increase (Scheuneman & Oakland, in press).

Conclusion

This paper has attempted to distinguish among legal and regulatory mandates, ethical issues, and professional responsibilities all which concern the appropriate use of tests and test data. Numerous efforts have been undertaken by testing professionals and professional organizations to improve responsible use of tests, yet often these efforts are judged to have fallen short. As tests are used by an increasing number of users with a variety of objectives (e. g., policy makers, state and local education officials, business) the potential for misuse of tests increases and efforts to educate and monitor test users become less effective. Existing testing standards and specialty guidelines and other forms of addressing the responsible use of tests are discussed. The potential consequences of testing and assessment are reviewed and a taxonomy has been proposed to aid test users in addressing the multiple and multidimensional consequences resulting from test use with various key stakeholder groups. Finally, this paper provides a more detailed review of the professional concerns arising from the migration of tests to a computer-based platforms and the increased demands placed on assessments in U. S. education.

The value of assessment is often related to its impact. Individual appraisals should bring to bear all relevant information to describe and explain important qualities, minimize problems, promote growth and development, and increase the validity of important decisions (e. g., course placement, admissions, certification, selection) (Scheuneman & Oakland, in press). National, state, and local testing programs should provide comprehensive data that can supplement other sources of information in both informing us of student skills and knowledge today and the growth in learning over time.

Legal, ethical, and professional concerns with assessment are difficult to distinguish. All such issues concern the proper use of assessment and the probable consequences of using assessments. Consequences of testing are in the eye of the beholder. The same assessment which presents several potential benefits to some groups (e. g., policy makers, community, business) may also result in negative consequences to individuals (e. g., test takers, students). A paradigm is needed to assist test users identify and evaluate the potential consequences that result from test use and the consequences which would result from alternative practices (use of more subjective processes, collecting no data). Additional attention to the consequences of testing, and how these are determined and evaluated by the various stakeholders is essential to reduce the misuse of testing and improve assessment practices among the increasingly diverse types of individuals using tests and results from testing.

Résumé

Les tests et les évaluations ont fait l'objet de contraintes supplémentaires au fur et à mesure que leur utilisation a été étendue par les enseignants, les employeurs et les décideurs. Des exigences grandissantes sont souvent adressées à ces mêmes évaluations par diverses instances, pour répondre à des buts variés (par ex: la réforme de l'enseignement, la responsabilité des étudiants, la qualité des méthodes d'enseignement). Ces tendances ont suscité des préoccupations considérables quant à l'emploi approprié des tests et de leurs résultats, et les méthodes de testing sont examinées de plus en plus scrupuleusement dans le domaine de l'éducation, de l'emploi et de la santé. Le présent article différencie les problèmes légaux, éthiques et professionnels apparus récemment du fait de demandes accrues d'évaluations et il identifie les problèmes spécifiques liés à l'application et à l'interprétation informatisées des tests. L'auteur passe en revue les efforts entrepris par les associations professionnelles en vue d'améliorer les pratiques d'évaluation ainsi que les problèmes concernant l'utilisation adéquate des évaluations dans le domaine de l'éducation. Enfin il propose une méthodologie pour identifier les conséquences liées à l'emploi des tests et une taxonomie destinée à évaluer les conséquences multidimensionnelles de l'emploi des tests dans un contexte donné.

Author's address:

Dr. Wayne J. Camara
The College Board
19 Hawthorne Drive
Princeton Junction, NJ 08550
USA
E-mail: wcamara@collegeboard.org

References

Aiken, L., West, S. G., Sechrest, L., Reno, Raymond R., Roediger III, H. L., Scarr, S., Kazdin, A. E., & Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North American. *American Psychologist, 45,* 721–734.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: APA.

American Educational Research Association & National Council on Measurements Used in Education. (1955). *Technical Recommendations for Achievement Tests.* Washington, DC: National Educational Association.

American Federation of Teachers, National Council on Measurement in Education, & the National Educational Association (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: Authors.

American Psychological Association (1953). *Ethical standards for psychologists*. Washington, DC: Author.

American Psychological Association (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.

American Psychological Association (1993). Ethical principals of psychologists and code of conduct. *American Psychologist, 49*, 1597–1611.

Association of Personnel Test Publishers (1990). *Model guidelines for preemployment integrity testing programs.* Washington, DC: Author.

Berliner, D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myths fraud and the attack on America's public schools.* Reading, MA: Addison-Wesley.

Brown, D. C. (February 8, 1997). Personal correspondence.

Camara, W. J. (1994). *Consequences of test use: The need for criteria.* Paper presented at the 23rd International Congress of Applied Psychology. Madrid, Spain.

Camara, W. J. (1996). Fairness and public policy in employment testing. In R. Barrett (Ed.) *Fair employment strategies in human resource management* (pp. 3–11). Westport, CT: Quorum Books.

Camara, W. J., & Brown, D. C. (1995). Educational and employment testing: Changing concepts in measurement and policy. *Educational Measurement: Issues and Practice, 14,* 1–8.

Center for Research on Evaluation, Standards and Student Testing (1995). *Results from the 1995 CRESST Conference: Assessment at the crossroads.* Los Angeles: UCLA, CRESST.

College Entrance Examination Board (1988). *Guidelines on the uses of College Board test scores and related data.* New York: Author.

Cole, N., & Willingham, W. (1997). *Gender and fair assessment.* Hillsdale, NJ: Erlbaum.

Council on Licensure, Enforcement, and Regulation (1993). *Development, administration, scoring, and reporting of credentialing examinations*. Lexington, KY: Council of State Governments.

Council on Licensure, Enforcement, and Regulation & National Organization for Competency Assurance. (1993). *Principles for fairness: An examining guide for credentialing boards.* Lexington, KY: Author.

Cronbach, L. J. (1988). Five perspectives on validity arguments. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Deutsch, C. H. (October 16, 1988). A mania for testing spells money. *New York Times*.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289–303.

Educational Testing Service (October 31, 1996). Test cheating scheme used encoded pencils, complaints charges. *ETS Access*. Princeton, NJ: Author.

Employee Polygraph Protection Act of 1988, Sec. 200001 et sec., 29 U.S.C.

Everson, H. E. (in press). A theory-based framework for future college admissions tests. In S. Messick (Ed.), *Assessment in higher education*. Hillsdale, NJ: Erlbaum.

Eyde, L. D., & Kowal, D. M. (1987). Computerized test interpretation services: Ethical and professional concerns regarding U. S. producers and users. *Applied Psychology: An International Review, 36*, 401–417.

Eyde, L. D., & Quaintance, M. K. (1988). Ethical issues and cases in the practice of personnel psychology. *Professional psychology: Research and Practice, 19*(2), 148–154.

Fairtest (Summer, 1996). Cheating cases reveal testing mania. *Fairtest Examiner, 9*, 3–4.

Hambleton, R. K. (1994). Guidelines for adapting psychological and educational tests: A progress report. *European Journal of Psychological Assessment, 10*, 229–244.

Haney, W. (1996). *Standards, schmandards: The need for bringing test standards to bear on assessment practice.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Haney W., & Madaus, G. C. (1991). In R. K. Hambleton & J. C. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 395–424). Boston, MA: Kluwer.

Haney, W., Madaus, G. C., & Lyons, R. (1993). *The fractured marketplace for standardized testing*. Boston, MA: Kluwer.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47*, 1083–1101.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education*. Washington, DC: Author. (Copies may be obtained from NCME, Washington, DC)

Joint Committee on Testing Practices. (1996). *Rights and responsibilities of test takers* (Draft). Washington, DC: Author.

Kirst, M. W., & Mazzeo, C. (1996). The rise and fall of state assessment in California 1993–96. *Kappan, 22,* 319–323.

Lindsay, G. (1996). Ethics and a changing society. *European Psychologist, 1*, 85–88.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational evaluation and policy analyses, 15*(1), 1–16.

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

Madaus, G. F. (1992). An independent auditing mechanism for testing. *Educational Measurement: Issues and Practice, 11*, 26–31.

Matarazzo, J. D. (1986). Computerized clinical psychological test interpretations: Unvalidated plus all mean and no sigma. *American Psychologist, 41,* 14–24.

McGrew, K. S., Thurlow, M. L., & Spiegel, A. N. (1993). An investigation of the exclusion of students with disabilities in national data collection programs. *Educational Evaluation and Policy Analyses, 15*, 339–352.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 33–46). New York: Macmillan.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9,* 287–304.

National Academy of Education (1992). *Assessing student achievement in the states: The first report of the National Academy of Education panel on the evaluation of the NAEP trial state assessment; 1990 Trail State Assessment.* Stanford, CA: Stanford University, National Academy of Education.

National Association of Collegiate Admissions Counselors (1995). *NCACA commission on the role of standardized testing in college admissions*. Author.

New York Times (January 22, 1997). One of the newest take-at-home tests: IQ. *New York Times.*

Paris, S. G., Lawton, T. A., Turner, J. C., & Roth, J. L. (1991). A developmental perspective of standardized achievement testing. *Educational Researcher, 20*(5), 12–20.

Phillips, S.E. (1996). Legal defensibility of standards: Issues and policy perspectives. *Educational Measurement: Issues and Practice, 15,* 5–13.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Educational Policy Analyses Archives, 5*(3) 1–18.

Sackett, P. R., Burris, L. R., & Callahan, C. (1989). Integrity testing for personnel selection: An update. *Personnel Psychology, 42,* 491–529.

Sackett, P. R., & Harris, M. M. (1984). Honesty testing for personnel selection: A review and critique. *Personnel Psychology, 32,* 487–506.

Scheuneman, J. D., & Oakland, T. (in press). High stakes testing in education.

Schmeiser, C. B. (1992). Ethical codes in the professions. *Educational Measurement: Issues and Practice, 11*(3), 5–11.

Society for Industrial and Organizational Psychology (1987). *Principles for the validation and use of personnel selection procedures.* Bowling Green, OH: Author.

Vroom, V. H. (1964). *Work and motivation.* New York: Wiley.