Consequences of Test Interpretation and Use: The Fusion of Validity and Values in Psychological Assessment

Samuel Messick

November 1988
Abstract

This paper addresses the role of social values in educational and psychological measurement, with special attention to the consequences of testing as validity evidence. Test use depends for its legitimacy on adherence to established principles of reliability, validity, and fairness. These principles are organized within an overarching conception of construct validity. Desired reliability depends on the meaning of the construct being measured while test fairness depends on comparable construct validity across groups or settings. In the name of fairness, value judgements need to be made throughout the testing process, as they do in the name of validity throughout the validation process.

Key Words: construct validity, test fairness, social values, test interpretation, test use, testing
This paper addresses the role of social values in educational and psychological measurement, with special attention to the consequences of testing as validity evidence, which is an inherently value-dependent enterprise. The primary measurement standards that must be met to legitimize a proposed test use are those of reliability, validity, and fairness, which are also value-laden concepts. Evidence of reliability signifies that something is being measured; the major concern is score consistency or stability. Evidence of validity circumscribes the nature of that something; the major concern is score meaning. Evidence of fairness indicates that score meaning does not differ consequentially across individuals, groups, or settings; the major concern is comparability.

The appropriate level of reliability depends on the meaning of the construct being measured because some constructs are conceived theoretically to be more consistent or stable than others. Hence, evidence of reliability consistent with the construct’s meaning is simultaneously also evidence of construct validity. Within these limits, the measurement intent is to achieve sufficient score consistency or stability to warrant the use of the scores in decision making. Another way of putting it is that the uncertainty involved in determining score levels should be inconsequential for the proposed test use.

Validity is mainly concerned with the meaning and consequences of measurement. Accordingly, validation studies aim to accrue convergent evidence supportive of score meaning and its action implications as well as discriminant evidence discounting plausible rival interpretations. Validity is a unitary concept, which means that fundamentally there is only one kind of validity, namely, construct validity. However, several complementary forms of evidence need to be integrated in construct validation,
evidence bearing on test content, score structure, substantive processes, generalizability, external relationships, and testing consequences.

In general, fairness implies impartiality, and the question arises as to how fairness is manifested in educational and psychological measurement. In particular, the impartiality entailed in test fairness is achieved through comparable construct validity across individuals, groups, and settings. That is, score levels should have the same meaning and consequences in different population groups and environmental contexts. This does not imply that fair test use yields equal group outcomes, however, because fair tests may validly document unequal outcomes resulting from, among other things, unequal opportunities to learn as well as differential experiences in learning and development.

The discussion thus far implies that both reliability and fairness are actually different subtopics subsumed under the overarching concept of validity. Indeed, reliability is treated here in terms of the generalizability aspect of construct validity. In contrast, fairness is discussed as a topic in its own right in order to highlight the complexity of the value choices and consequences entailed in fairness judgments in relation to validity judgments.

Thus, by virtue of focusing on the role of values in educational and psychological measurement, this paper emphasizes the value-laden nature of validity and fairness as psychometric concepts. However, validity and fairness are not just psychometric principles. They are themselves social values that have meaning and force outside of measurement whenever judgments and decisions are made. We must keep in mind this broader meaning because it informs the public’s mind as to what is valid and fair. Ultimately, our psychometric considerations need to be justified in this public arena.
The Values of Validity and Fairness

Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment (Messick, 1989). Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. Hence, what is to be validated is not the test or observation device per se but rather the inferences derived from test scores or other indicators (Cronbach, 1971) -- inferences about score meaning or interpretation and about the implications for action that the interpretation entails. Broadly speaking, then, validity is an inductive summary of both the existing evidence for and the potential consequences of test interpretation and use.

To validate an interpretive inference is to ascertain the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported. To validate an action inference requires not only evidence of score meaning but also justification of value implications and action outcomes, especially appraisals of the relevance and utility of the test scores for particular applied purposes and of the social consequences of using the test scores for applied decision making. Thus the key issues of test validity are the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use.

Although different sources and mixes of evidence are required to support particular score-based inferences, validity is a unitary concept in the sense that score meaning as embodied in construct validity underlies all score-based inferences. Hence, all validation
is construct validation. Furthermore, because value implications as a basis for action are
integral to score meaning, construct validation has to include an appraisal of these value
implications in terms of the actual and potential consequences of test use.

This point was forcefully made by Cronbach (1988) in his treatment of validation as
persuasive argument: “the argument must link concepts, evidence, social and personal
consequences, and values” (p.4). A special concern is to guard against adverse
consequences that are traceable to sources of invalidity such as construct
underrepresentation or construct-irrelevant variance (Messick, 1989, 1995). Indeed, it is
difficult even to isolate questions of the value implications of score interpretations from
questions of the validity of those interpretations. Even the meanings of the words “valid”
and “value” derive from the same Latin root, valere, meaning “to be strong.” From this
perspective, validity judgments are value judgments.

This fusion of validity and values in test interpretation and use is especially salient in
connection with testing consequences. This is the case because consequences are
evaluated in a validity arena where social values abound, namely, in terms of test
fairness. As we have seen, test fairness is a validity issue because fairness implies
comparable validity across individuals, groups, and settings (Willingham, 1998;
Willingham & Cole, 1997). The key problem in test fairness is legitimizing the meaning
of the term “comparable” in terms of value choices.

The remainder of this paper explores the role of values in test validity and fairness in
more detail. First, six main aspects of the validation process are examined, many of
which entail important value issues and value choices. Next, we consider some of the
value issues that need to be addressed in establishing the fairness of the test development
process as well as the fairness of test use. Finally, the controversy over consequences as validity evidence is confronted, as is the value-dependence of such evidence.

Validity of Test and Use

In essence, test validation is empirical evaluation of the meaning and consequences of measurement, taking into account extraneous factors in the applied setting that might erode or promote the validity of local score interpretation and use. Because score meaning is a construction that makes theoretical sense out of both the performance underlying the score and its pattern of relationships with other variables, the psychometric literature views the fundamental issue as construct validity.

Recurrent Questions of Validity

To evaluate the meaning and consequences of measurement is no small order, however, and requires attention to a number of persistent validity questions, such as:

- Are we looking at the right things in the right balance?
- Does the assessment adequately cover the content domain? Has anything important been left out?
- Does our way of looking, especially the choice of stimulus and response formats, facilitate the expression of competence or other processes, or are sources of invalidity or irrelevant variance introduced that bias the scores or judgments?
- Is the scoring system consistent with the structure of the domain about
which inferences are to be drawn or predictions made?

- Are there plausible rival interpretations of score meaning or alternative implications for action and, if so, by what evidence and arguments are they discounted?

- Are the scores reliable and are their properties and relationships generalizable across the contents and contexts of use as well as across pertinent population groups?

- Are the value implications of score interpretations empirically grounded in terms of observed outcomes, especially if pejorative in tone, and are they commensurate with the score's trait implications?

- Do the scores have utility for the proposed purposes in the applied setting?

- Are the scores applied fairly for these purposes, that is, consistently and equitably across individuals and groups?

- Are the short- and long-term consequences of score interpretation and use supportive of the general testing aims and are there any adverse side-effects traceable to test invalidity?

The general thrust of such questions is to seek evidence and arguments to discount the two major threats to construct validity -- namely, construct underrepresentation and construct-irrelevant variance – as well as to evaluate the action implications of score meaning.
Constructing Construct Validity

Validity is now widely viewed as an integral or unified concept (APA, 1985). Therefore, establishing validity requires the collection and integration of multiple complementary forms of evidence to answer an interdependent set of questions such as those just considered. To make this explicit, it is illuminating to differentiate unified validity into several distinct aspects to underscore issues and nuances that might otherwise be downplayed or overlooked.

In particular, six distinguishable aspects of construct validity are highlighted: content, substantive processes, score structure, generalizability, external relationships, and testing consequences. In effect, these six aspects function as general validity criteria or standards for all educational and psychological measurement (Messick, 1989, 1995).

These six aspects of construct validity take the construct to be measured as a given. However, there is a critical prior step in test design, if not in test validation, that must be carefully negotiated because it is profoundly value-laden. This is the stage of construct choice for the particular testing purpose. Although this choice is greatly influenced by salient variables operative in the criterion domain, there is usually ample leeway for social values to intrude. Is it not at least in part a value decision when reading and mathematics are privileged over other abilities in high-school graduation examinations and in college admissions (Willingham & Cole, 1997) or when intelligence test scores are used in placing children in special education programs (Heller, Holtzman, & Messick, 1982)? Approaches to circumscribing the nature of the construct being measured are addressed by the six aspects of construct validity that we turn to next. Although these considerations may not determine construct choice, they can certainly modify the nature
of that choice. Especially pertinent in this regard are the content, substantive, and structural aspects of construct validity.

The *content* aspect of construct validity (Lennon, 1956; Messick, 1989) includes evidence of content relevance and representativeness as well as of technical quality (e.g., appropriate reading level and unambiguous phrasing). Key issues here are specifying the boundaries of the content domain to be assessed and selecting tasks or stimulus questions that are representative so that all important parts of the domain are covered. The aim is to assemble a representative sample of tasks or items tapping knowledge, skills, or other attributes relevant to the purpose of the testing. Expert professional judgment is usually relied upon to document content relevance and representativeness.

The *substantive* aspect of construct validity refers to the psychological processes (of thought, motive, or feeling) underlying both task or item performance and the correlations across items. There are two important points: One is the need for tests that assess domain *processes* in addition to traditional coverage of domain *content*; the other is the need to move beyond traditional professional judgment of content to accrue empirical evidence that the assumed cognitive, conative, or affective processes are actually at work. Thus, the substantive aspect adds to the content aspect of construct validity the need for empirical evidence of response consistencies or performance regularities reflective of domain processes (Embreton, 1983; Loevinger, 1957; Messick, 1989). In one form or another, the substantive aspect calls for models of the cognitive processes required by the assessment tasks.

The *structural* aspect appraises the degree to which the score scales are consistent with the structure of the domain being measured. The internal structure of the assessment
should reflect the internal structure of the construct domain. Scoring models should be rationally consistent with what is known about structural relations inherent in behavioral manifestations of the construct in question (Loevinger, 1957; Peak, 1953). The theory of the construct domain should guide not only the creation of assessment tasks but also the development of scoring criteria and rubrics.

The *generalizability* aspect examines the extent to which scores and interpretations are consistent across assessment tasks, populations, and settings (Cook & Campbell, 1979; Feldt & Brennan, 1989; Shulman, 1970). An assessment should provide representative coverage of the content and processes of the construct domain being tested, so that the score interpretation not be limited to the particular sample of assessed tasks but be generalizable to the construct domain more broadly. Evidence of such generalizability depends on the degree of correlation of the assessed tasks with other tasks representing the construct or aspects of the construct. This issue of generalizability of score inferences across tasks and contexts goes to the very heart of score meaning. Indeed, setting the boundaries of score meaning is precisely what generalizability evidence is meant to address.

In one sense, the generalizability aspect of construct validity intersects with reliability in that it refers to the consistency of performance across the tasks, occasions, and raters of a particular assessment (Feldt & Brennan, 1989). But in a second sense, generalizability refers to *transfer* or the consistency of performance across tasks or items that are representative of the broader construct domain. Transfer refers to the range of tasks that performance on the assessed tasks facilitates the learning of or, more generally, is predictive of (Ferguson, 1956). Phrased more broadly to make clear that personality as
well as ability tests are encompassed, transfer refers to the range of behaviors that the assessed behaviors facilitate the learning or acquisition of or are predictive of.

The issue of generalizability is especially relevant to performance assessments, which are typically limited to a small number of tasks because each performance task requires an extensive amount of time. There is a conflict in performance assessment between time-intensive depth of examination on any one task and the number of tasks needed for representative domain coverage. This conflict must be carefully negotiated in designing and interpreting performance assessments (Wiggins, 1993).

The external aspect of construct validity refers to the extent to which performance on a test is related to other test and nontest behaviors, the correlations to be either high or low as predicted by the theory of the construct being assessed. Thus, the meaning of the scores can be substantiated externally by appraising the degree to which empirical relationships with other measures, or the lack thereof, are consistent with that meaning. That is, the constructs represented in the assessment should rationally account for the external pattern of correlations.

It is especially important to examine the external relationships between the test scores and criterion measures of the desired behaviors the test is intended to predict when using the scores for selection, placement, licensure, certification of competence, program evaluation, or other accountability purposes in applied settings. Once again, the construct theory points to the relevance of potential relationships between the test scores and criterion measures, and empirical evidence of such links attests to the utility of the scores for the applied purpose.
The consequential aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short and long term. Ideally, there should be no adverse consequences associated with bias in scoring and interpretation, with unfairness in test use, or with negative effects on teaching and learning.

Testmakers are mainly concerned with minimizing negative impact on individuals and groups due to any source of test invalidity such as construct underrepresentation or construct-irrelevant variance (Messick, 1989). That is, validity is compromised when the assessment is missing something relevant to the focal construct that, if present, would have permitted the affected examinees to display their competence. Similarly, scores may be invalidly low because the measurement contains something irrelevant that interferes with the affected examinees' demonstration of competence. In contrast, adverse consequences associated with valid measures of an individual’s current status -- such as validly low scores resulting from poor teaching or limited opportunity to learn -- are not the testmakers' responsibility but the test users'. Adverse consequences resulting from valid assessment are problems not of measurement but of social or educational policy.

**Validation As Evidence-Based Argument**

From the discussion thus far, it should be clear that test validity cannot rely on any one of these complementary forms of evidence. Neither does validity require any one form, or a high level of every form, if there is good overall evidence supporting score meaning. What is required is a compelling argument that the available evidence justifies the test interpretation and use, even though some pertinent evidence may be lacking.
The six aspects of construct validity apply to all educational and psychological measurement, including performance assessments or other alternative assessment modes. Taken together, they provide a way of addressing the multiple and interrelated validity questions that need to be answered in justifying score interpretation and use. In constructing the validity argument, one can set priorities about the forms of evidence needed to justify the inferences drawn from test scores by focussing on the vulnerable points in the argument that require support (Kane, 1992; Shepard, 1993). The key point is that the six aspects of construct validity provide a means of checking that the rationale or argument linking the evidence to the inferences drawn touches the important bases. If not, an argument should be provided that explains why such omissions are defensible (Messick, 1995).

**Fair Assessment and Equitable Treatment**

There remains an important, overarching issue in the validity of test use, namely, test fairness. A central question in test validity asks whether any adverse consequences derive from sources of test invalidity such as construct underrepresentation or construct-irrelevant variance, which bears directly on the appropriateness and acceptability of outcomes. A central question in test fairness asks whether the consequences of test use are appropriate and acceptable, which relates directly to the validity of the inferences drawn from the test scores. Thus, the concept of fairness in assessment is impossible to divorce from the concept of validity because the two share a mutuality of meaning and import.
Fairness, like validity, is not just a psychometric issue. It is a social value, and there are alternative points of views about its essential features. In essence, fairness implies impartiality, with an absence of prejudice or favoritism. In regard to test use, this impartiality derives from comparable construct validity. A fair test is one that yields comparably valid scores from person to person, group to group, and setting to setting (Willingham, 1998). So, for instance, if an assessment yields scores that consistently underestimate ability levels for members of a particular group, then the test would be considered unfair to that group. Or if an assessment is interpreted to measure a single construct across groups, but in actuality measures different constructs in different groups, it would also be unfair.

**Alternative Views of Fairness: Process or Outcome**

Although there are alternative views as to what constitutes test fairness, most of them relate back in one way or another to the central idea of comparable validity. For instance, the 1998 draft revision of the *Standards for Educational and Psychological Testing* cites four alternative views of fairness commonly found in the technical and popular literature. Two of these views characterize test fairness, respectively, as the absence of bias and as equitable treatment of all examinees in the testing process.

Bias arises when deficiencies in the test itself result in different meanings for scores earned by different subgroups. An example is item content that differentially reflects the cultural or geographical background of different examinees or groups. Fairness as equitable treatment of all examinees in the testing process requires that examinees have a comparable opportunity to demonstrate their competence or, more generally, their
standing on the construct being measured. Fair treatment includes such factors as comparable testing conditions and equal access to test familiarization and practice materials. There is general agreement that tests should be free from bias and that all examinees should be treated equitably in the testing process. These first two views refer to test fairness in ways that stop just short of requiring comparable validity across individuals, groups, and contexts, but they are quite compatible with the notion of comparable validity.

The other two views characterize test fairness, respectively, as equal opportunity to learn and as equality of testing outcomes. Opportunity to learn is an important issue to take into account when evaluating the comparability of score meaning across groups and is especially problematic when the scores are used to make high-stakes decisions about individuals. However, it is important to note that comparable score meaning does not necessarily imply the same or identical score meaning, as will be seen shortly in connection with efforts to accommodate tests to the needs of examinees with disabilities or those with limited English proficiency.

From one perspective, the notion that test fairness requires equality of testing outcomes is given some credence on the grounds that equality of outcomes should be the ultimate indicator of genuine equality of learning opportunities as well as of genuine equity of treatment. However, given current educational and social realities, the idea that test fairness requires overall passing rates to be comparable across groups is not generally accepted in the professional literature.

Indeed, the second two views refer not so much to fairness in testing as they do to equity of social and educational experience (Gordon, 1998). It is important not to
confound test fairness with educational or social equity because the two need to be pursued in their own right and in different ways. Unequal outcomes and opportunities do not in themselves signify test unfairness because tests may validly document unequal individual or group outcomes that are reflective in part of unequal opportunity to learn.

It is also important to distinguish between equality (the state of being the same) and equity (treatment that is just under the circumstances) and to recognize that not all inequalities are inequities. Indeed, in education as in medicine, the watchword should not be equal treatment but, rather, treatment that is appropriate to the characteristic and sufficient to the need (Gordon, 1998). From this perspective, it was important that the Standards phrased one of the alternate views of test fairness just considered as equitable, not equal, treatment of all examinees in the testing process, thereby allowing room for accommodations to the different needs of examinees with handicaps or allowing use of tests translated into an examinee’s best language (Willingham et al., 1988).

The Values of Accommodations in Testing

The issue of equity and the attendant desire to accommodate the testing process to the different needs of examinees with disabilities or those with limited English proficiency is a prime example of the role of social values in educational and psychological measurement. The pursuit of equitable testing accommodations goes to the very heart of comparable construct validity and test fairness.

In this regard, it is important to distinguish two kinds of comparability (Willingham et al., 1988). One kind, called score comparability, means that the properties of scores (such as reliabilities, internal relationships among items, and external relationships with other
variables) are comparable across groups and settings. Score comparability is important in justifying uniform score interpretation and use for different groups and in different circumstances. The other kind, called task comparability, means that the test tasks elicit comparable cognitive processes across different groups and different settings.

Within task comparability, two types of cognitive processes may be distinguished: those that are relevant to the construct measured and those that are ancillary to the construct but nonetheless involved in task performance (Wiley, 1991). Comparability of construct-relevant processes is needed to sustain common score meaning across groups and contexts. Again, these processes need not be identical from person to person or from time to time for the same person, as long as it can be maintained that they represent alternative or substitutable construct-relevant processes, such as when different strategies are employed in a test of reasoning or problem solving.

Ancillary processes, in contrast, may in many cases be deliberately modified for some individuals or groups without jeopardizing score meaning. Such modifications, when they are possible, provide a fair and legitimate basis for accommodating tests to the needs of examinees with disabilities or those with limited English proficiency (Willingham et al., 1988). For example, a fair accommodation might be to read a mathematics test out loud to an examinee with visual impairment because reading is ancillary to the construct being measured. In this regard, the availability of multimedia stimulus and response modes on the computer promises generalized accommodation in computer-based testing to the various needs of examinees with disabilities (Bennett, 1998).

Thus, comparable validity—and hence, test fairness—does not require identical task conditions but, rather, comparable construct-relevant processes with ignorable construct-
irrelevant or ancillary processes that may be different across individuals and groups. Such accommodations, of course, have to be justified with evidence that score meaning and properties have not been unduly eroded in the process.

**Fairness Issues Throughout the Testing Process**

Fairness, like validity, cannot be properly addressed as an afterthought once the test has been developed, administered, and used. Fairness issues, like those of validity, must be confronted throughout the interconnected phases of the testing process, from test design and development, through administration and scoring, to interpretation and use (Willingham, 1998). Indeed, one of the most critical fairness issues occurs at the design stage, namely, the choice of the constructs to measure for the particular testing purpose. For example, consider the possible test requirements for awarding high-school diplomas. Graduation rates for males and females, as well as for language minority students, would be quite different if the tests emphasized reading and writing as opposed to science and mathematics (Willingham & Cole, 1997). Any limited number of subjects covered by the test would likely yield differential graduation rates for different groups because they underrepresent the broad construct of school learning and because students have differential opportunities to learn. Some alternatives are to assess school learning more comprehensively, to use more than one assessment mode such as high-school grades as well as test scores, and to justify any limited choice of subjects in terms of the social values of the school and community.

Construct choice is a critical issue for both fairness and validity because that choice determines test content, appropriate test formats, and test specifications. Ultimately,
construct choice influences not only test development but also administration, scoring, interpretation, and use. Furthermore, the particular test use, in the first place, is critical in determining construct choice. Thus, these several decisions are interconnected, but the fairness issues involved at each stage are different and require different approaches.

There are other fairness considerations at the stage of test design and development. For example, the format of the items or the content used to contextualize an item may be more familiar to some examinees and groups than to others and may thereby create a disadvantage for some. Such potential bias in test questions is usually addressed empirically by examining whether individuals who have the same level of ability, but are members of different groups, have the same probability of getting the items correct.

Fairness issues arise in the administration of tests because of nonstandard testing conditions. Examples include the use of test forms of unequal difficulty and providing a less-than-optimal testing environment (in terms of lighting, space, and temperature) that may disadvantage some examinees more than others. Fairness is also an issue whenever scoring is not completely objective, because subjective scoring is routinely threatened by biases associated with construct-irrelevant characteristics of the performance, the examinee, and the setting.

Fairness issues in connection with test use abound. Some examples are using measures that are less valid than others available, relying unduly on a single score, and basing decisions on an underrepresented view of the relevant construct when such uses result in differential individual or group impact (Willingham, 1998). A widespread and important use of tests is for selection, as in college admissions or job entry. Considerable attention has been given in the past to fair selection models, which by their very nature put heavy
emphasis on comparable predictive validity but at the expense of comparable construct validity. Because these models expose alternative views of fairness as well as the specter of conflicting social values in the pursuit of fairness from different perspectives, we examine them next in more detail.

In the late 1960s and early 1970s, measurement specialists attempted to define fair selection in psychometric terms (Cleary, 1968; Cole, 1973; Linn, 1973; Thorndike, 1971). In one way or another, all of these fair selection models attempted to cope with the fact that any imperfect predictor fails to select members of a low scoring group in proportion to their criterion success rate. This is the case even when individuals having the same predictor score are equally likely to succeed regardless of group membership—that is, even when the predictor yields a common regression line for the different groups, which is Cleary’s (1968) formulation of an unbiased predictor. Because fair test use implies that selection decisions will be equally appropriate in some sense regardless of an individual’s group membership and because different selection systems yield different proportions of selected individuals in population groups displaying different test performance levels, questions of test fairness arise in earnest in selection settings. The problem for each of the fair selection models is to specify the meaning of the phrase in some sense.

Cleary’s (1968) sense of fairness, as was just seen, is that individuals with the same predictor score should be equally likely to succeed regardless of group membership. From another viewpoint, Cole’s (1973) conditional probability model of fair selection holds that applicants who can achieve a satisfactory criterion level should have an equal probability of being selected regardless of group membership. Several other models of
fair selection were formulated and contrasted (e.g., Darlington, 1971; Einhorn & Bass, 1971; Linn, 1973, 1976; Thorndike, 1971), each of which assumed, either explicitly or tacitly, an unbiased criterion measure that is uniformly and fairly applied across groups. This unbiased criterion then serves as the standard for evaluating potential bias in the predictor measures.

Some of these models proved to be mutually incompatible and even contradictory (Petersen & Novick, 1976). It soon became apparent that each model accords a different importance or social value to the various quadrants of selected versus rejected and successful versus unsuccessful individuals in the different population groups (Dunnette & Borman, 1979). For example, one could emphasize equality in the rates of false negatives across groups, that is, equality in the rate of rejecting individuals who would have succeeded if selected. Moreover, the values accorded could be a function not only of desired criterion performance but also of desired individual and group attributes (Novick & Ellis, 1977). As a consequence, each model not only constitutes a different definition of fairness but also implies a particular ethical position (Hunter & Schmidt, 1976). Each view is ostensibly fair under certain conditions, so that arguments over the fairness of test use turn out in many instances to be disagreements as to what the conditions are or ought to be (Messick, 1989).

The core social problem in selection is the need to reduce subgroup differences in selection rate while minimizing any negative effects this might have on overall predictive validity. In an attempt to achieve such a tradeoff, a National Research Council committee recommended the use of subgroup norming by means of within-group percentiles or some such device (Hartigan & Wigdor, 1989). The rationale is in tune with Cole’s (1973)
conditional probability model, but as an action proposal tantamount to the adjustment of subgroup scores it precipitated vigorous controversy (e.g., Gottfredson, 1994; Sackett & Wilk, 1994). In any event, the argument became moot when Congress expressly prohibited any form of score adjustment on the basis of race, color, religion, sex, or national origin (Civil Rights Act of 1991). In light of this Congressional prohibition, we must return to other views of fair selection.

With the recognition that fundamental value differences are at issue in fair selection, several utility models were developed that require specific value positions to be articulated (e.g., Cronbach, 1976; Gross & Su, 1975; Petersen & Novick, 1976; Sawyer, Cole, & Cole, 1976), thereby incorporating social values explicitly with measurement technology. However, the need to make values explicit does not determine or make easier the hard choices among them. At this point, it appears difficult if not impossible to be fair to individuals in terms of equity, to groups in terms of parity or the avoidance of adverse impact, to institutions in terms of efficiency, and to society in terms of costs and benefits all at the same time (Messick, 1989).

In principle, a workable balancing of the needs of each of the parties is likely to require successive approximations over time, with iterative modifications of selection procedures based on their consequences to date, so that the interests of none of the parties gets too far out of line. In practice, however, such balancing of needs and values almost always comes down to a political resolution. The challenge is that by recognizing the role of social values in educational and psychological measurement, we might be able to inform these political choices with evidence about effective policy alternatives and their consequences.
In any event, predictive validity, regardless of the selection model on which it is based, is not sufficient to signify test fairness. What needs to be comparable across groups and settings for fair test use is score meaning and the actions based thereon. That is, test fairness derives from comparable construct validity, which subsumes but goes beyond evidence of comparable predictive validity.

Controversies Over Consequences

The elevation of construct validity to an overall validity framework for evaluating test interpretation and use has been highly controversial, especially with respect to the role of testing consequences as validity evidence contributing to score meaning. Critics argue vigorously that testing consequences should not be included as an integral part of validity. This is partly due to their misapprehension that the emphasis on adverse consequences traceable to sources of invalidity primarily reflects a concern with test misuse. For example, Wiley (1991) maintains that “the understanding of these use errors is conceptually and socially important, but involves social and moral analyses beyond the scope of test validation…and would needlessly complicate the conception and definition of test validity” (p.89). Popham (1997) believes that one of the motives for including consequences as validity evidence “was surely to draw our attention to the unsound uses of test results” (p. 12).

On the contrary, the motivation was to draw attention to unanticipated side-effects of legitimate test use, especially if unanticipated adverse effects are traceable to sources of invalidity such as construct underrepresentation and construct-irrelevant variance. Of
course, procedural errors and unsound interpretations would invalidate a particular local test use, and such *misuse* should not contribute to the validation process. By restricting their understanding of the consequential basis of construct validity to the effects of test misuse, these critics find it easy to argue that the validity of score inferences is separable from and orthogonal to the consequences of test misuse (Mehrens, 1997; Popham, 1997; Tenopyr, 1996). The argument is easy because it is basically true but immaterial. However, it is not possible to separate score meaning from the consequences of legitimate test interpretation and use. The unanticipated consequences of legitimate score interpretation and use bear not only on the justification of the use but also on the soundness of score meaning and, hence, are an integral part of the validation process (Messick, 1998).

Other critics (e.g., Brandon, 1996) contend that the consequential basis of construct validity is either circular or redundant, or both. They agree that adverse consequences signal invalidity only if traceable to sources of invalidity such as construct underrepresentation or construct-irrelevant variance. However, they argue that these sources of invalidity are identified and taken into account using the methods associated with Messick’s (1995) other five validity aspects bearing on content, substantive processes, score structure, generalizability, and external relationships. Therefore, for these critics, consequences are unnecessary for examining validity.

However, even after diligently applying the methods of the other five validity aspects, every test still underrepresents its construct to some degree and contains sources of irrelevant variance, if for no other reason than it is a test and not a criterion performance (Loevinger, 1957). Test makers usually maintain that this remaining invalidity is
inconsequential until confronted with evidence to the contrary. This is precisely why unanticipated consequences constitute an important form of validity evidence. Unanticipated consequences signal that we may have been incomplete or off-target in test development and, hence, in test interpretation and test use (Messick, 1998).

Testing consequences are inherently value-laden because their import is contingent on the value system brought to bear in evaluating them. Hence, the consequential basis of validity is value-dependent, a complexity that leads critics to reject it and to turn to the evidential basis in the hope that it is value-independent (Markus, 1998). However, value-neutral as well as theory-neutral facts are problematic in the postmodern world, so the evidential basis of validity can hardly be considered independent of values (Messick, 1989). But neither is the consequential basis independent of evidence.

This interplay of facts and values is one reason that the validation process needs to be rhetorical as well as scientific. In the process of validation, one needs to craft a persuasive argument linking constructs, evidence, and values (Cronbach, 1988), which is the rhetorical part of the enterprise. The scientific part relies on convergent and discriminant evidence involving multiple measures and multiple methods. In any event, the operation of social values in test design, development, interpretation, and use cannot be avoided but must be carefully negotiated in testing practice in the cause of both validity and fairness.

Indeed, there are signs that the measurement field is moving beyond the question of whether or not testing consequences should be part of validity (Yen, 1998). Instead, questions are being asked about how such consequences might be examined and who is responsible for collecting the evidence about consequences (Linn, 1998).
Overview

In sum, test use depends for its legitimacy on adherence to established principles of reliability, validity, and fairness. These principles are organized within an overarching conception of construct validity. The form and level of desired reliability depends on the meaning of the construct being measured while test fairness depends on comparable construct validity across groups and settings (Messick, 1989; Willingham, 1998). Conceptions of test fairness entailing only selected aspects of construct validity, such as the fair selection models based on comparable predictive validity, are not sufficient in themselves to justify fair test use. What is needed are comparable score meaning across groups and settings as well as comparable test uses based on the action implications of the construct meaning. That is, the purpose of the testing and the construct measured should mutually support or justify each other, in the sense that the testing purpose implies constructs to be measured and the construct choice implies testing purposes.

Pursuing test fairness as comparable construct validity requires a heavy burden of convergent and discriminant evidence bearing on test content, score structure, substantive processes, generalizability, external relationships, and testing consequences (Messick, 1995). From a practical standpoint, it would of course be more convenient and expedient to be able to base test fairness on only a selected aspect of construct validity such as comparable prediction across groups and settings. Apart from the insufficiency of the argument, however, there is also the danger that if evidence of comparable prediction were not forthcoming, the prize of fairness would be lost (possibly for the wrong reasons). In contrast, the pursuit of comparable construct validity is more flexible. With
multiple complementary forms of evidence to be addressed, one does not have to rely on any one of them, nor is any one of them required. That is, just as no single form of evidence—whether content-based, criterion-related, or consequential—is sufficient to justify test interpretation and use, neither is any particular form of evidence necessary. What is necessary is compelling rationale justifying score meaning and its action implications from the mosaic of available evidence, even though some pertinent evidence had to be forgone.

Fairness issues arise in each stage of the testing process from design and development, through administration and scoring, to interpretation and use. Thus, in the name of fairness, value judgments need to be made throughout the testing process, as they do in the name of validity throughout the validation process, especially in regard to testing consequences. The pervasive role of social values in educational and psychological measurement is not intrusive but rather integral to the assessment process because, in the last analysis, fairness judgments and validity judgments are value judgments.
REFERENCES


annual meeting of the Society for Industrial and Organizational Psychology, San Diego.


Authors Note

This paper was presented at the International Congress of Applied Psychology,