Measure, Mismeasure, or Not Measurement at All?
Psychometrics as Political Theory

Mark J. Garrison, *D'Youville College*

Abstract

The author of this article challenges a common assumption made by both critics and defenders of standardized-testing technology (or psychometry), namely that standardized tests "measure" something (culture, ability, etc.). It argues that psychometric practice cannot be classified as a form of measurement and instead is best understood as a marker of social value, an inherently political act. The chapter concludes by suggesting the significance of this argument for debates regarding "standards," "accountability," and educational assessment more generally.

Introduction

As the role and significance of standardized testing and thus psychometrics increases with new state and federal laws mandating yearly high stakes tests in the name of accountability and a high quality education for all, debates about the validity of this technology have intensified in academic as well as lay circles. Yet, an assumption typical of both critics and supporters of standardized testing is that such tests *measure something*. This belief is evidenced by the frequently heard question: what is being *measured* by an I.Q. or a standardized achievement test? Standardized tests are developed, supporters argue, to measure a student's overall academic ability or potential. They can measure the degree to which any student has mastered a particular body of knowledge or set of skills, what is given as scholastic achievement. Proponents of standards-based reform argue that these tests are well suited to measure the performance of educational systems, in both relative

and absolute terms, and identify areas in need of improvement. In this way, proponents of the use of standardized tests to improve education are in favor of the recent federal education act known as *No Child Left Behind*.

Yet, the federal mandates of this Act for schools that do not meet expected performance levels—including the withholding of federal monies and the imposition of various restructuring schemes such as charter schools, for-profit school management, and state takeovers of entire districts—are considered punishments by many individuals, especially given the perennial critiques of standardized tests as such.

For example, one can easily find critics arguing that the best predictor of a student's or school's score on any given standardized test is their associated social class—what one commentator called the Volvo Effect: family wealth as indicated by brand of car predicts student test performance (Wesson, 2000). The format and content of these tests, critics continue, reveal profound cultural and gender biases as well. Such examinations measure more of a student's cultural familiarity with the discourse of Western schooling and linguistic socialization and less some objective notion of innate intelligence or even what they know and are able to do. While critics often mount strong arguments that standardized tests are indeed biased, their assumption of measurement typically holds.[1]

In this way, Gould's (1981) now classic book on intelligence testing, *The Mismeasure of Man* has popularized a critique of testing predicated on the assumption that a key problem lies with the *misuse* of educational and psychological tests and specious interpretations of the *meaning* of test scores. The problem is not with the technology itself, but simply how it is used.[2] This is in fact consistent with the path psychometry has followed. Now fundamental psychometric concepts such as construct validity have emerged to articulate meaningfulness in relation to how the test is used (Cronbach & Meehl, 1955; Thorndike, 1971).

In this article, I take a different direction. I argue that psychometry fails to meet its claim of *measurement* and that its object is not the measurement of non-physical human attributes, but the marking of some human beings as having more worth and value than other human beings, an act central to maintaining the legitimacy of a particular kind of hierarchical social system such as a meritocracy. Psychometry's claim to measurement serves to veil the fundamentally political act of marking social value.

Definitions of Psychometry

According to the Oxford English Dictionary, psychometry literally means measuring the soul, or "mind measuring." The first reported use of the word, appearing in 1854, gave psychometry as the "faculty of divining, from physical contact or proximity only, the qualities or properties of an object, or of persons or things that have been in contact with it." Sense two is given as follows: "The measurement of the duration and intensity of mental states or processes" with the

following quote from Francis Galton. "Psychometry . . . means the art of imposing measurement and number upon operations of the mind, as in the practice of determining the reaction-time of different persons." (Galton's choice of the word *imposing* should not go unnoticed.) And finally, the Oxford English Dictionary offers this definition for psychometrics, which has a more contemporary flare: "The science of measuring mental capacities and processes; the application of methods of measurement to the various branches of psychology."

The literal, etymological meaning of psychometry is a useful place to begin. What would it mean to *measure mind* (let alone soul)? Is mind (or soul for that matter) the kind of thing one has more or less of? Or, to start with the more contemporary definition, are mental capacities such that they exist in gradation? Is a theory of mind needed in order to determine if mind can be measured, and if so, how metrication can take place? The issues raised here are fundamental from the point of view of both the theory and practice of measurement, and addressing them serves as a useful starting point for deliberating on the nature of measurement and the status of psychometry as a science.

The Scientific Status of Psychometry

Measurement deals with the dialectical relationship between quantity and quality (Berka, 1983; Nash, 1990). Psychometry, however, renders measurement as the mere application of number systems to objects, a process that, according to Cicourel (1964) has "no necessary reference to the empirical world" (p. 8). Thus a key problem for psychometry rests with its adoption of S. S. Steven's definition of measurement as the allocation of numerals according to rules, which stands as the basis for psychometric practice (Lorge, 1951). Such a definition causes some critics to describe psychometry as "measurement by fiat" (Block & Dworkin, 1976; Cicourel, 1964; Pawson, 1989).

According to Berka (1983), scales producing a mere linear ordering are not measurements proper—if this were the case, almost anything of this nature (the assignment of numbers to a given phenomenon) could be considered measurement. For example, a questionnaire, in which the respondent expresses his or her attitude with the aid of numbers, cannot be taken as an instance of *measurement* of preferences" (Berka, 1970, p. 145). Nash (1990), elaborates the problems inherent in this approach:

> It is easy to construct a series of questions and treat the resulting form as a scale. Consider, for example, these items, to be rated "agree," "uncertain," "disagree": "I get on well at school," "School is a neat place," "My teachers often praise my work." So, what do these questions "measure"? This is by no means irrelevant to practice—it is the very question which validity questions are designed to answer. Such items, as anyone with a little knowledge of conventional psychological research will recognize, might easily turn up in a scale designed as an Academic Self-Concept

> Instrument and they might just as equally turn up on a Pupil School Evaluation Instrument. Furthermore, it is more likely than not that a psychologist employing both "measures" would never so much as look at the test items, but report that the study found Academic Self-Concept . . . to be highly correlated with Pupil School Evaluation and offer that as evidence for the conclusion that pupils with high academic self-concept are also satisfied with their school. (p. 132)

Psychometry is plagued by the tendency to imbue data with properties of the measurement procedure. Coombs (as cited in Cicourel, 1964) points out that one cannot assume a scale to be a property of the "measurand" if it (that scale) is a necessary consequence of the method of analysis (p. 13). The relevance here to educational testing is striking. It is not permissible to argue that intelligence (or any purported characteristic of individuals or groups) is normally distributed in the population on the basis of the normal distribution of scores, for such a distribution is demanded by the tests most commonly used. While "such a procedure ought to be virtually standard practice by the lights of many current parametric techniques," Pawson (1986) writes, "to do so . . . is a 'gratuitous' expression of statistical expedience" (p. 56).

Central to Berka's (1983) theory of measurement is the concept of *magnitude*, defined as the property of relative size or extent: it can be thought of as *how much*. Common, everyday conceptions such as length and weight represent known magnitudes. A standard (say the meter) must be theoretically and technically fit for the measure of objectively existing properties of a thing or phenomenon—it is what allows for a single property, again, say length, of different objects to be measured. It is also the standard that allows for equivalence, or what is also conceptualized as calibration. It is important to understand that while a standard is necessary for measurement, at least initial theoretical work is presupposed for it to be able to render magnitudes. For example, there needs to be a conception of the qualitative aspect of heat before its measurement can take place (see Block & Dworkin, 1976; Pawson, 1989). Once such theoretical knowledge is at least initially established measurement may be possible, allowing for the location and study of the relation between quantitative and qualitative change in objects, phenomena, and processes. For Berka (1983), the issue is not primarily one of precision of measurement. Metrication includes the claim that laws and rules governing quantitative and qualitative change are being represented mathematically.

Theoretical work also determines if the property or quality under investigation *can* be measured. The development of measurement has generally progressed from classification (qualities), to topology (comparisons) to metrication (measurements) (Berka, 1983). Classification concepts such as "cold" become topological when comparisons are used, such as colder *than* . . . . Thus they "enable us, not only to establish the sameness (or difference), but also to mutually compare at least two objects which possess a given property and, consequently, to arrange them into a sequence" (Berka, 1983, p. 6).

Topological concepts provide a transition from classificational to metrical

concepts according to Berka (1983), who argues that classification (or differentiation) itself is not measurement, contrary to what some contemporary textbooks put forward (e.g., Hopkins, Stanley & Hopkins, 1990, pp. 1-3). A key problem, one pointed to above with the definition of psychometry as *mind measurement*, is the assumption that the mind or a purported function of mind is a property capable of gradation. There are many properties that do not permit gradation—such as *Pilsner*, *feline*, *wooden*, and *human*. In other words, the psychometric dictum of E. L. Thorndike that if something exists, it must exist in some amount is false.

Yet, this premise may have great social and political significance. We know that humanity is riddled with cases of some human beings being designated as less than human, or not human at all; *humanness* has been given as something individual persons and groups "have more or less of"—a key presupposition of the eugenicist's project of a "master race." The U.S. Constitution made this presumption when it rendered African slaves and Native peoples as only holding a fraction of the value of white Europeans. Such designations are based on the claim that some human beings have less intelligence, ability, or otherwise valued attribute, than other human beings, and on that basis, they have been rendered less human, of less value, and in some cases, a threat to civilization itself.

Most readers will accept at some level, however, that education is something that can be graded. Clearly some students learn more of a particular subject matter than others, and clearly people obtain, both officially and in practice, different levels of education in different fields, etc. In this way, measuring educational achievement seems less problematic than measuring mind or intelligence. Defining content areas such as math seem relatively simple.

Yet, the project of measuring academic knowledge in practice appears particularly fixated on ranking *human beings* and less on achievement *per se*. For example, norm-referenced achievement tests offer results in terms of percentile ranks, not delineations of what a student does or does not know about a given field of study, let alone diagnoses of the cause of difficulty. Put another way, scoring in the 70th percentile only indicates how well one did relative to the norm; it does not indicate 70 percent of required material was mastered. Thus the test remains at the *topological level*, where percentile results indicate only that, for example, Sue performed *better than* Joe; the preceding semantics suggest that the object is in fact the ranking of persons, and not what they know or can do as such.

The same problem exists with so-called measures of ability. Nash (1990) contends that norm-referenced ability tests only provide rank order information. "Students are ranked, in effect, by their ability to correctly answer test items, but it is inaccurate to argue that their 'cognitive ability' is therefore being measured" (Nash, 1990, p. 63).

State tests marking schools and districts as performing at the proficient level in a given area also reflect this problem—these levels or benchmarks cannot be shown to correspond in a consistent and sufficiently precise way to an actual body

of knowledge or set of skills (e.g., Haney, 2000). Because these tests take as their object the differentiation and ranking of human persons, and do not reflect how much has actually been learned in school by any given set of students, such tests should not be used for accountability purposes (Popham, 1999; Wesson, 2000). Ranking human worth on the basis of how well one competes in academic contests, with the effect that high ranks are associated with privilege, status and power, suggests that psychometry is premised on Anglo-American political ideals of rule by *the best and the brightest*, a *meritocracy*.

Psychometricians' Consciousness of the Problem

In the following cases, testing experts admit that their tests are not measurements, or indeed, do not measure any real entity. Yet, as soon as these problems are identified, they are ignored or dismissed on the basis of being practical.

None other than Alfred Binet admitted that his intelligence test did not constitute a measurement of intelligence, while simultaneously going on to elaborate his scale as a measuring instrument.

> The scale properly speaking does not permit the measure of the intelligence, because intellectual qualities are not superposable, and therefore cannot be measured as linear surfaces are measured, but are on the contrary, a classification, a hierarchy among diverse intelligences; and for the necessities of practice this classification is equivalent to a measure. (Binet & Simon, 1916, p. 41)[3]

This "intellectual bad faith," as Nash (1990) calls it, is also evident among current practitioners. Under a section entitled "Limitations of Achievement Tests" Ebel and Frisbe (1991) make these frank admissions:

> Unlike the inch or pound, the units used in measuring this ability cannot be shown to be equal. The zero point on the ability scale is not clearly defined. Because of these limitations, some of the things we often do [always do, in reality] with test scores, such as finding means and standard deviations, and correlation coefficients, ought not to be done if strict mathematical logic holds sway. Nonetheless, we often find it practically useful to do them. When strict logic conflicts with practical utility, it is the utility that usually wins, as it probably should. (p. 31)

Suen (1990) writes that the "ability of test scores to truthfully reflect quantities of a characteristic of interest actually involves a huge inferential leap" (p. 5). Steyer (1989) begins his work by noting the conclusion of Suppes and Zinnes (1963) that "psychological tests are 'pseudopointer instruments,' the readings of which 'do not correspond to any *known* fundamental or derived numerical assignment'" (p. 26). To take one more example, Von Broembsen, Gray, and Williams (1974) lament the fact that few scientists are willing to undertake theoretical work in measurement. They regrettably admit that there is "no simple correspondence (or isomorphism) between mathematical structures and structures of relations between elements in the social sciences as there are, for example, in mathematical and physical sciences" (pp.

51-52). Yet, in a footnote on the next page, the authors say this problem is given too much attention.

For a measurement system to be valid there must be a correspondence between elements, relations and operations of the mathematical and substantive system in question. This raises the criteria of measurement, or the problem of being *isomorphic*—a correspondence that allows for, as an example, the additive principle (one can take 10 feet and add it to 10 feet and obtain 20 feet) and other mathematical procedures. Test items are not equivalent in the manner degrees are on a thermometer, for example, although it must be noted that the additive principle does not hold when dealing with temperature, yet there is a systematic correspondence between the readings from a thermometer and heat phenomena. Thus, as there is no proof to being isomorphic with the object of measurement, let alone a clearly specified and theorized object of measurement, educational and psychological tests cannot claim to reflect laws and rules governing psychological processes. Nash (1990) concludes that the "necessary conditions for metrication do not exist" (p. 145).

Validity

In traditional psychometric theory, validity is defined as the degree to which a test measures what it claims to measure. I want to first point out the oddity of this formulation. For example, how does the reader respond to this: my ruler is valid to the degree to which it measures length? Is it normal practice to begin ruler validation by asking this seemingly circular question? Rulers by definition measure length. Note as well that by asking what a test *measures* the assumption that something *is* being measured goes unchallenged.

In sorting this out the importance of theory is revealed for in reality a ruler is a *standard of length*, and can be my forearm, the meter, the inch. My forearm does not measure length any less or more than the inch—both can represent the length of an object.[4] Just to make the point clear, my forearm may be longer or shorter than the reader's, yet both can measure that property called length.

There are generally four types of test validation: predictive, concurrent, content and construct. Nash (1990) has rendered a powerful critique of the psychometric discourse of validity, while others have analyzed the persistent confusion between facts and values (Block & Dowrkin, 1976; Schiff & Lewontin, 1986). However, for the purposes here, I want to highlight validity discourse as a form of justification.

According to Wilson (1998), quoting the American Psychological Association (1985) standards, "Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores." It goes on immediately to explain that: "Test validation is the process of accumulating evidence to support such inferences" (chap. 16, ¶5). Two things stand out. First, how is it possible that the meaning of any test lies somewhere outside the test? Nash (1990) writes: "the view that

meaning lies anywhere other than in the test-text must be resisted" (p. 133). If the test-text does not contain the meaning of the test—many students and teachers feel such tests are indeed meaningless—then it is not possible for the test taker to grasp the meaning of the test. Intervention of specially trained judges (experts), who will decide the meaning of the test for the taker and the consumer of that information, is thus required and justified. This brings us to the second point, which is that "accumulating evidence to support such inferences" is unabashedly what Wilson (1998) calls *advocacy*. He explains further: "The 1985 Guidelines describes an ideal validation as including several types of evidence." "However," the APA manual emphasizes, "the quality of the evidence is of primary importance, and a single line of solid evidence is preferable to numerous lines of evidence of questionable validity" (Wilson, 1998, p. 9). Wilson rightly points to the "tautology and redundancy in the phrase questionable validity" as "remarkably inept." But even more significant is that,

> validity is proposed as the characteristic of the evidence used to support the construct "validity," and the essence of the concept is surely its very questionability. Far more damning, however, is the clear implication that evidence that does not cogently support the assertions of the test users should not be presented. Putting it another way, validity is a concept based on advocacy, is a rationalizating tool for a methodological decision already made, and is an ideological support rather than a scientific enterprise. (Wilson, 1998, chap. 16, ¶ 14-15)

Assessment and Social Value

The above analysis suggests that standardized tests are not tools in measurement. Here I suggest that they be explored as assessments, and in particular as tools for the marking of social value. Williams (1990) specifies that social value is "used in the sense in which value is socially attached to groups as well as to structural positions via status duality (good or bad) and spatial duality (high or low)." This duality of "the sacred and the profane" thus constitutes "the two levels of value duality within a system of vertical classification," or ranked categories (p. 1). With these propositions from Williams, I am immediately struck how educational testing is an elegant example of vertical classification. Within such a system all individuals and groups "are then placed in either the sacred or the profane position. Theoretically, they are mutually exclusive categories." With this perspective, "vertical classification is conceived of as the rigid segregation of human beings into categories of good or bad and high or low" (Williams, 1990, p. 1). It seems to me that marking virtue (good or bad) and talent (high or low) constitute the object of standardized test based assessment within a hierarchically structured social system. Because students do not typically come to school with official labels, academic achievement and ability must be constantly assessed. Herein lies one general reason for the ubiquity of testing, as well as the basis for testing being equated with opportunity.

While it is common to suggest that measurement is simply a more precise form of assessment (e.g., Hopkins, Stanley & Hopkins, 1990, pp. 1-3) assessment is a distinct undertaking. Wiggins' (1993) remarks are useful here.

> *Assess* is a form of the Latin verb *assidere*, to "sit with." In an assessment, one "sits with" the learner…. The person who "sits with you" is someone who "assigns value"—the "assessor" (hence the earliest and still current meaning of the word, which relates to tax assessors). But interestingly enough, there is an intriguing alternative meaning to that word, as we discover in the Oxford English Dictionary: this person who "sits beside" is one who "shares another's rank or dignity" and who is "skilled to advise on technical points." (p. 14)

Wiggins suggests that there are hierarchical and non-hierarchical forms of assessment, variously signified in the notion that assessment can be done *to* a person, or *with* that person of the *same rank*. Thus he senses the importance of *power relations* as a feature of assessment. As Wiggins laments, it is the hierarchical form which dominates the history of assessment in education.

But significance also rests in the definition of assessment as the judgment of value. This is the distinguishing feature of standardized test-based assessment, and it is precisely this fact that is vigorously avoided with psychometric pretenses of measurement.

Measurement and Assessment

Standards are the foundation of both assessment and measurement. In measurement, the objective of the standard is magnitude, the abstract expression of objectively existing qualities of things and phenomenon.[5] The object of the standard in assessment is value; the *relation* here is between subject and object. With measurement, the magnitude makes possible the grasping of the *relation* between quality and quantity. Standards in assessment make possible the judgment of value by stipulating boundary points as indicators of quality (merit, worth, goodness, authenticity). In fact, official educational assessment operates on the basis of establishing desired qualities and their vertical classification, or placement in vertically structured category systems with the assistance of numbers. This is what is being delineated when it is said that the task of validity is to determine the meaning (value) of test scores. The validity discourse about test score meaning relative to testing purpose is based on value not residing in things or phenomenon themselves, *but in their relation to subjects*. Length, however, is a property of an object.

The confusion between measurement and assessment is not insignificant, having both scientific and ideological importance. Scientifically, the confusion over what is measurement is bound up with confusing properties of objects with properties of numbers (a good example being the normal curve) and social relations and the properties of those objects or phenomena in the relation. These mistakes function to masking the workings of the values system in official testing practices and the power involved in making value judgments about people.

Assessment, Value, Position

It appears that assessment—the use of standards in the judgment of value—is a feature of the earliest forms of stratified human society. Ever since "the initiation ceremonies of early societies, there have always been arrangements for formally recognizing the capacity to perform important and recognizable social roles and to exercise the associated social status and power" (Eggleston, 1986, p. 59). A role's associated status and power reveals, first, its importance or value. Second, there are really two capacities being referred to—the first is the capacity to perform the role itself, and the second capacity is to exercise the role's *associated social status and power*. The second capacity or ability can be thought of as a quality of the individual or the social status of the individual. It is this second ability, which I think is the ultimate object of assessment via standardized tests in education. This may very well explain the relatively strong correlation of test scores and SES (note the uncritical replication of social value in the notion socio-economic status) compared to the relatively week correlation of ability and achievement tests with actual performance (e.g., see National Commission on Testing and Public Policy, 1990).

Precision as Marker of Value

The more advanced the standard, the more that thing or phenomenon is valued. Kula (1986) observes, "in societies where land was relatively abundant, the system of area measures tended to be poorly developed" (p. 6), and the contrary in societies where land was scarce. The same tendency is observed with measures of weight. With "the Ashanti of Ghana," one finds a very advanced system of weights, "in whose economy the extraction of gold dust played a major part" (Kula, 1986, p. 6). Speaking specifically about the question of value, Kula argues that, "the more valuable the object, the finer the measure [standard] employed in its measurement" (p. 88). He continues noting that determinations of measuring procedures are partly practical, but cautions, "practical considerations afford only a partial explanation . . . . Our emotive, 'feeling' attitude to the object centers upon its 'value' for man" (p. 88).[6]

This general proposition can be seen at work with social value if we take the example of driving a truck versus becoming a physician. Driving a truck can be said, on this basis, not to be of great social value, for the standard to obtain such a license is not very fine, or precise, even though the safety of millions of travelers and billions of dollars worth of products are at stake. One either passes or fails the relevant tests; as with SAT and ACT scores no elaborate hierarchy of licenses exists, and the tremendous difference in the value of degrees from different academic institutions. Academic achievement and ability, the standards for entering medical school, are thus highly valued, reflected on the fineness of their measure. The great effort towards precision is not based on measurement, but instead constitutes a means by which to identify and produce value. That there is a great deal of fineness in the standard of the second and not in the first suggests which is held in more esteem by the dominant culture. We might expect this situation to change radically if truck drivers somehow

got themselves involved in making transportation policy. That is, those who are deemed to occupy sacred positions (good character, high ability) are fit to make decisions, to decide on the all important questions of who, what, where and when.

Within the theory of a natural aristocracy or what is commonly call meritocracy there is the assumption that talent signifies virtue, and that virtue qualifies someone for decision-making positions. As with the early Chinese examinations, while character was the preferred object of examination, *ability was substituted as the measure (read: standard) of man* (Menzel, 1963). Within the Enlightenment tradition of the West, ability is viewed as a marker of virtue. For example, a high score on an IQ test suggests a student is worthy of being trained to play social roles with high status and power—the high score suggests the high status, worth or virtue. School exams focus on abstruse academic exercises, I think, because they endeavor to assess the ability to exercise a role's attending social status and power—e.g., is he or she capable of "good judgment"—and not so much the functional capacities demanded by the role. That is, official educational assessments seem primarily concerned with the second capacity identified above. Because the role of truck driver currently has little associated status or power, licensure procedures need only focus on the functional ability itself.

Exercising status and power demands a particular set of aims and values, or else the stability of that status and power is threatened (it is this stability of the status quo which seems to be the referent of "good judgment"). Abstruse academic exercises constitute values; their exercises constitute what is valued, reflecting a definite world outlook. For example, within Euro-American thought, written competitive exams reveal a person's ability to delay gratification, or "self-denial." Roach (1971) argues that reformers in England

> put considerable stress on the moral argument at both the individual and the national level. For the individual, examinations are a test of common-sense and of character as well as of book-learning. To do well in them demands perseverance and self-denial which strengthen the character. For the nation, a competitive system would be based on high moral principle and would help to reduce corruption and place seeking. (p. 30)

That is, the quality that is reflected in doing well on tests of "common sense" and "book learning" is "good character." It is character that is the object of the examinations. By the middle of the 19th Century, on both sides of the Atlantic, written competitive exams pointed to these general abilities, more towards reasoning and less on the memorization of facts. In British context, this ability reflected the gentlemanly values of *quick wit* as well as the ability to reason or judge well. The Cambridge Tripos, a test in logic and reason, rewarded speed and accuracy, just as these qualities insured success in the courtroom (Roach, 1971, pp. 13-14). In fact, much of this discourse is found in *A Nation at Risk* with its talk of "excellence" and "commitment to a set of values" as the basis of "the learning community" (National Commission on Excellence in Education, 1983).

These notions of ability, of capacity, are bound up with social roles, for ability

must have a place for it to be manifest. This *quality or state of being able* manifests itself in the "physical, mental, or legal power to perform," according to Webster's. Note that ability can both signify a power inhering in persons—such as physical power or mental power—as well as legal power, or being formally allowed to do something. It is in the context of the present society that mental power stands as one justification for legal power. It is significant, I think, that the etymology of ability is from the Middle English, suitability. In this regard, standardized test-based assessment is the judgment of worth relative to a structural slot or social position—what is deemed of value and who is deemed of value—the meeting place of which is variously achievement or ability. That is, achievement and ability signify both places and persons, as in someone (an individual) who becomes rich (a social position). Note as well that suitability can take individuals or positions as its object—is the individual suitable to the position; is the position suitable for the individual.

Summary and Implications

The long-standing debate as to whether standardized tests accurately measure *merit* (worth) is simultaneously a frank admission that standardized tests aim to assign value to human beings—to determine who is worthy of what type of education—and a block to grasping fully the significance and implications of such a project.[7] Standardized tests are not designed to accurately and fairly select, certify, and monitor via measurement of specific competencies or abilities, but rather to legitimate such acts via the *assessment of social value*. Thus it may be more useful in analyzing psychometry to view it as political theory, as a formal justification for a system where "the argument for democracy is not that it gives power to men without distinction, but that it gives greater freedom for ability and character to attain power" (E. L. Thorndike quoted in Karier, 1973, p. 122). It is no wonder that results obtained by these methods closely parallel the inequalities upon which the entire economic and political order is based.

Possibly the first implication of this understanding is to reject any form of assessment that functions to differentially *value* human persons. Let me be clear: the issue is not in recognizing that humans differ in their abilities, interests and so on (though such difference are not, in my view, the problem they can be made out to be). The problem emerges when such differentiation is systematically linked to a hierarchical social structure and the reproduction of that structure.

Thus there is a need for assessment in education to establish a new starting point, one predicated on the equal worth, dignity and rights of human beings and human cultures.[8] Those working to develop assessments in the service of education must vociferously reject the linking of academic prowess with notions of bad or good. The habit of talking of *good students* must be replaced with a culture where the work of teachers, students and the community as a whole is judged by teachers, students and the community as a whole on the basis of whether or not this collective work is serving

to prepare youth to solve the problems they and their society face. This is the basis upon which assessments should take place, and in fact such a drive may under pin recent efforts towards alternative or authentic assessment, in particular those predicated on Gardner's (1993) notion of multiple intelligences, which opens up space to recognize and value a broad range of diverse human abilities and achievements.

This conclusion has profound implications for the present standards and accountability movement, especially as embodied in the *No Child Left Behind Act 2001* (NCLB). It suggests to me that strategies opposing *NCLB* on the basis that it does not provide enough funds to meet legal requirements misses the fact that *NCLB*, and in particular its testing mandates, are in themselves attacks on public education and those who attend and work in public schools. Based on what has been presented here, the law's functioning to mark so many of the nation's schools as failures is not an aberration. By marking public schools as failures, this standards based reform is in my view devaluing public education and education itself. It is an effort, among other things, to assimilate Americans to a lower standard of education, not to a higher one.[9] It is a clear message by those in power that the arrangement whereby all are to have the quintessential American opportunity of education is being wrecked, to be replaced with the notion that only those who "perform well" *deserve* an education (as opposed to it being a right).

The words measure, measures or measurement appear, by my count, 135 times throughout the nearly 700-page *NCLB* law, giving the reform initiative the appearance of being politically neutral or objective. In fact, claims to measurement appear to be part of an effort to render decisions about who gets what, when, where and how—i.e., political decisions—as apolitical, as not needing the full participation of those whom they affect. For example, as one textbook puts it: "…assessment, measurement, and evaluation are in the best interests of students…and society—the educational decisions are only as good as the quality of the information on which they are based" (Hopkins, Stanley, & Hopkins, 1990, p. 3). *Good decisions* are one's that are based on *good quality* measurements and this stands as a justification for the fact that these decisions ultimately serve one set of interests over another, often those of ruling elites over those of students, teachers and parents.

Notes

[1] See Haney (1993) for a discussion of the notion of bias in psychometry. For a broad sociological discussion of the relationship between social structure and forms and purposes of assessment, see Broadfoot (1979; 1996). See Madaus (1994) for an overview of equity issues in test-based reform.

[2] See Madaus (1990) for an interesting discussion of testing as a social technology.

[3] Note Gould (1981) quotes this same passage (p. 151) but does not include the clauses beginning with "but" where Binet equates classification with measurement.

[4] Note that the precision of these standards is quite a different matter, and here the issue of standardization also arises; however these issues will not be dealt with for space considerations.

[5] Many ask whether magnitudes exist prior to measurement, or whether they are the result of measurement. For Berka (1970, pp. 147-149) "magnitudes are, in substance, of a relational nature. They depend on the procedure of measurement and the resultant numerical values. From this it does not … follow that magnitudes do not have an objective basis in reality." Measurement *is* dependent upon the activity of humans and their standards, but it does not follow from this that the object of measurement is a product of human will.

[6] The utility of applying Kula's (1986) study of physical measures to the study of extra physical measures such as educational and psychological tests is supported by Garrison's (2001) case studies of the emergence of standardized testing.

[7] Examples of this language can readily be found on Fairtest's website (www.fairtest.com). For example, a quick search turns up "Test Scores Are Not 'Merit,'" as a response to the attack on minority college students in the name of high standards.

[8] For those wanting to explore the relevance of human rights for education in general, see for example Spring (2001). A useful discussion of rights and opportunities as they relate to education can be found in Esconi and Hurwitz (1974). For the impact of test-based reform on equity issues, see Madaus (1994). Also see the Right to Education Project (RTE) the only such web site in the world devoted solely to the right to education (http://www.right-to-education.org). It was started by Katarina Tomasevski, the first ever Special Rapporteur on the Right to Education of the United Nations Commission on Human Rights, after her appointment in 1998.

[9] I have made this conclusion on the basis of the large body of research and commentaries showing that the testing policies embodied in *NCLB* actually serve to lower the level of education in general, and exacerbate already gross inequalities. For example, see: (Apple, 2001; Horn & Kincheloe, 2001; McNeil, 2000; Orfield & Kornhaber, 2001).

References

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

Apple, M. W. (2001). *Educating the "right" way: Markets, standards, god, and inequality*. New York: Routledge.

Berka, K. (1970). Methodological problems of measurement. *Problems of the Science of Science, 1*, 143-153.

Berka, K. (1983). *Measurement: Its concepts, theories, and problems* (A. Riska, Trans.). Boston: Kluwer.

Binet, A., & Simon, T. (1916). *The development of intelligence in children* (E. S. Kite, Trans.). Baltimore: Williams & Wilkins Company.

Block, N. J., & Dworkin, G. (1976). IQ, heritability, and inequality. In N. J. Block & G. Dworkin (Eds.), *The IQ Controversy* (pp. 410-540). New York: Pantheon.

Broadfoot, P. (1979). *Assessment, schools and society*. London, UK: Methuen.

Broadfoot, P. (1996). *Education, assessment and society.* Buckingham, UK: The Open University.

Cicourel, A. V. (1964). *Method and measurement in sociology*. London, UK: The Free Press of Glencoe.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. In A. W. Ward, H. W. Stoker, & M. Murray-Ward (Eds.), *Educational measurement: Origins, theories and explications* (Vol. 1, pp. 179-208). Lanham, MD: University Press of America.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.

Eggleston, J. (1986). Examining school examinations; A sociological commentary. *New Education, 8*(1), 59-69.

Esconi, C. A., & Hurwitz, E. (Eds.). (1974). *Education for whom? The question of equal educational opportunity*. New York: Dodd, Mead & Company.

Gardner, H. (1993). *Frames of mind: The theory of multiple intelligences* (2d ed.). New York: Basic Books.

Garrison, M. J. (2001). *Education, standards and the assessment of social value: Case studies in the origin of standardized testing*. Unpublished doctoral dissertation, University at Buffalo (SUNY).

Gould, S. J. (1981). *The mismeasure of man*. New York: W. W. Norton & Company.

Haney, W. (1993). Testing and minorities. In L. Weis & M. Fine (Eds.), *Beyond silenced voices: class, race and gender in United States schools* (pp. 45-74). Albany, NY: State University of New York Press.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved December 1, 2001, from http://epaa.asu.edu/epaa/v8n41/

Hopkins, K. D., Stanley J. C., & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation*. Englewood Cliffs, NJ: Prentice Hall.

Horn, R., & Kincheloe, J. (Eds.). (2001). *American standards: Quality education in a complex world, the Texas case*. New York: Peter Lang.

Karier, C. (1973). Testing for order and control in the corporate liberal state. In C. Karier, P. C. Violas, & J. Spring (Eds.), *Roots of crisis: American education in the twentieth century* (pp. 108-137). Chicago: Rand McNally & Company.

Kula, W. (1986). *Measures and men*. Princeton, NJ: Princeton University Press.

Lorge, I. (1951). The fundamental nature of measurement. In A. W. Ward, H. W. Stoker, & M. M. Ward (Eds.), *Educational measurement: Origins, theories and explications* (Vol. 1, pp. 11-37). New York: University Press of America.

Madaus, G. F. (1990). *Testing as a social technology*. Boston: Boston College.

Madaus, G. F. (1994). A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. *Harvard Educational Review, 64*(1), 76-95.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.

Menzel, J. M. (1963). *The Chinese civil service; career open to talent?* Boston: Heath.

Nash, R. (1990). *Intelligence and realism: A materialist critique of IQ*. New York: St. Martin's.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Retrieved April 4, 2001, from http://www.ed.gov/pubs/NatAtRisk/risk.html

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America*. Chestnut Hill: National Commission on Testing and Public Policy.

Orfield, G., & Kornhaber, M. L. (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: Century Foundation Press.

Pawson, R. (1986). On the Level: Measurement Scales and Sociological Theory. *Bulletin de M'ethodologie Sociologique, 11*(July), 49-82.

Pawson, R. (1989). *A measure for measures: A manifesto for empirical sociology*. London, UK: Routledge.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8-15.

Roach, J. (1971). *Public examinations in England 1850-1900*. London, UK: Cambridge University Press.

Schiff, M., & Lewontin, R. (1986). *Education and class: The irrelevance of IQ genetic studies*. Oxford, UK: Clarendon Press.

Spring, J. H. (2001). *Globalization and educational rights: An intercivilizational analysis.* Mahwah, NJ: Lawrence Erlbaum.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability and testability. *Methodika, 3*, 25-60.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 1-76).

Thorndike, R. L. (1971). Educational measurement for the seventies. In A. W. Ward, H. W. Stoker, & M. Murray-Ward (Eds.), *Educational measurement: Origins, theories and explications* (Vol. 1, pp. 87-106). Lanham, MD: University Press of America.

Von Broembsen, M. H., Gray, L. N., & Williams, J. S. (1974). Formalization and verification of theory in the behavioral sciences. *International Behavioral Scientist, 6*(4), 51-67.

Wesson, K. (2000, November 22). The "Volvo Effect." *Education Week, 20*(12), 34, 36-7.

Wiggins, G. (1993). *Assessing student performance*. San Francisco: Jossey-Bass.

Williams, R. (1990). *Hierarchical structures and social value: The creation of Black and Irish identities in the United States*. Cambridge, UK: Cambridge University Press.

Wilson, N. (1998, May 22). Educational standards and the problem of error. *Educational Policy Analysis Archives, 6*(10). Retrieved June 23, 1998, from http://epaa.asu.edu/epaa/v6n10/

About the Author

Mark Garrison is Assistant Professor of Education at D'Youville College, in Buffalo, New York. He received his doctorate in the social foundations of education from the University at Buffalo in 2001. His book *The Political Origins of Failure: Education, Standards and the Assessment of Social Value* is due out fall 2006 (from State University of New York Press). He also has forthcoming material on the social context of the use of educational technology.