The Gordon Commission
on the Future of Assessment in Education

Epistemology and Measurement: Paradigms and Practices

I. A Critical Perspective on the Sciences of Measurement

Ezekiel J. Dixon-Román
University of Pennsylvania

Kenneth J. Gergen
Swarthmore College

The political and economic interests of Western industrialist nations have centered on scientific measurement as the dominant orientation to effective policy making. The Federal turn to evidence-based policy and decision making represents a recent instantiation. This orientation has especially materialized in education policy, more generally, and its reliance on testing, in particular. The past ten years of Federal education policy have relied on standardized measures of student achievement as means to evaluate school performance. Policy makers have assumed that the instruments of measurement are sound, robust, and valid. At the same time, these assumptions rest on an early philosophy of science of measurement, a philosophy that has until recently been given little critical attention (Borsboom, 2008; Michell, 1999; Mislevy, 1997). In the two, yoked contributions that follow we shall offer both a critical analysis of the epistemological grounds of traditional measurement, and provide an alternative epistemology that, in our view, holds far more promise for educational practices and policies of assessment. It is in the context of the emerging transformations in world conditions that we view this shift from a positivist/empiricist to a social constructionist epistemology as paramount in significance.

The first of these two papers lays out the main tenets and assumptions of positivist epistemology, and discusses how the positivist paradigm became a philosophical movement ultimately employed in the legitimation of the social sciences. We will pay particular attention to the influence of positivism on measurement and how measurement became a hallmark Modernist science. We then shift our focus to the history and philosophy of science of measurement tracing each of its epistemological shifts and assumptions. Given that the major shifts in psychological measurement occurred with the development of measurement theory, this section will be organized around each of the major theoretical paradigms of measurement – i.e., classical test theory, latent variable modeling, and representational models. We also discuss the more recent shifts and emergent approaches to measurement that take more of a sociocultural and situative perspective.

In the final section, we raise and explore critical questions of the positivist paradigm of measurement that should be addressed as we move into the future. These questions include: Why should measurement sciences sustain the assumptions of universality at the cost of the multiplicity in social and cultural particularities? In what ways can the seeming objectivity represented in instruments of measurement be removed, so as to reveal the ideological investments that inhabit the production and reading of the instruments? And, is it still

theoretically tenable to assume that the manifest behavioral responses to items provide any sort of privileged access to the inner workings of the human mind? Through this line of critical questioning we set the stage for the considerations of alternative epistemological possibilities in the second paper, and their implications for practice in the broader domain of assessment to which measurement is a limited special case.

A word must be said about the multiple audiences we wish to reach with our proposals. As suggested above, issues of measurement, testing, assessment, evaluation, and educational policy making while not the same are closely linked. If, for example, one challenges the common assumption of validity in measurement, there are important ripple effects in terms of our understanding of testing, assessment, and evaluation. By the same token, with a different view of the educational process and educational policy, our orientation to testing, assessment, and evaluation would be different. While we recognize different forms of tests, assessments, and evaluations our concern here is more squarely with those forms that are more directly (and indirectly) influenced by the measurement sciences. There is no means by which we can effectively speak to all professional audiences at once. However, we shall endeavor to do the kind of bridge-work that will make the central arguments clear and ultimately relevant to future policies.

As a précis to the analyses that follow, it will first be useful to take brief account of what many see as seismic changes taking place in world conditions. Such changes amplify the weaknesses in the traditional paradigm of measurement and invite exploration into the implications of a paradigm more adequately attuned to the emerging social conditions.

Global Change: The Fluid and the Frozen

There is now voluminous commentary on what are commonly sensed as major changes taking place in cultural - and indeed global life. Although a vast simplification, for present purposes it is useful to see these changes as comprising two, inter-dependent movements, the one toward disorganization and upheaval, and the other toward systematization and control - or metaphorically, movements in the opposing directions of the *fluid* and the *frozen*. On the side of massive and disruptive change, for example, Marshall Berman (1988) focuses on the mounting ambiguities and complexities thrust upon the culture in the past century. Hardison (1995)

describes the general drift away from the vision of nature as solid and tangible, and the accompanying acceleration in social change. Ritzer (2009) ascribes major transformations in patterns of contemporary life to newly emerging channels of consumption. Eitzen and Zinn (2011) stress the effects of globalization on contemporary life. Rodgers (2011) describes the contemporary condition as an "age of fracture," focusing on the loss of collective purpose; Bauman (2011) views our condition as one of increasing "liquidity," with early cultural traditions replaced by the continuing demands for the new.

Yet, none of these works touch sufficiently on mammoth changes in communication resulting from a raft of newly emerging technologies. In the time it takes to read this sentence out loud, for example, over 80 million email messages will have been launched into the world. In the last year alone it is estimated that 8 trillion text messages were sent via cell phones. And this is to say nothing of the increasing reach of television, or the 1.5 billion people in the world who surf the web, or the estimated over 177 million tweets per day. This enormous expansion in communication not only facilitates rapid change, but it also multiplies the possibilities for organizing and stabilizing behavior. As people communicate together they create meaning; they generate together a sense of what is taking place, what is valuable to do, and appropriate logics of action. The existing technologies of communication enable people to locate like- minded others from across their societies, and indeed from around the world. Thus we find, for example, an enormous expansion in the number of religious sects, NGOs, on-line interest groups, and politically active enclaves, and the potential for their participants to remain in communication 24/7. In effect, there is a splintering of societies into multiple groupings, or as Maffesoli (1996) has put it, we live in a "time of tribes." At the same time, at every institutional level - in business, government, and religion, for example - this same potential for increased organization takes place. Rules, plans, requirements, surveillance and the like are all used to stabilize order. And, precisely these many and increasing forces for disruption bring about increased demands for control.

By and large, the dominant practices of measurement in education (as exemplified in standardized tests) have been employed in the service of freezing the social order. Typically they are used to establish and sustain standards, and to facilitate policies that ensure effective performance in these terms. As one might say, national testing in education is used to order the society in terms of the values and rationalities of those occupying positions of institutional

power. To be sure, there is much to be said for realizing a vision of society in which all members enjoy the benefits of education. However, in light of the emerging conditions of change, this vision becomes increasingly perilous. In particular, it does not take into account:

- The ability of multiple groups - religious, ethnic, political, and so on - to organize themselves around alternative visions of education, to develop agendas, and to pose a critical challenge to what appear as educational impositions from elsewhere.
- The continuously shifting character of what counts as useful knowledge. With rapid transformations in world conditions, new topics and requirements are constantly developing, rendering older curricula irrelevant. Demands for courses related to environmental sustainability, digital capability, the media, and the Middle East are illustrative. Differing groups in different parts of the nation will also view these educational needs in different ways.
- The continuously shifting character of pedagogy. Traditional top-down, and pre-formed pedagogical practices are becoming outmoded by virtue of the communication skills, proclivities, and needs of contemporary students. Teachers are less capable of creating and sustaining a "dominant reality" or set of values within their classes, as communication technologies enable "counter-realities" to rapidly form within student populations. In effect, what is taught and how it is taught become increasingly dependent on innovation and improvisation.

It is within this context that we now turn to the justifying logics of the sciences of measurement.

The Positivist Paradigm and the Sciences of Measurement

One of the earliest forces preparing the ground for contemporary measurement practices was the emergence of 19th century positivist philosophy. Particularly in the writings of Auguste Comte, positivist philosophy was an appropriation of the scientific method to the study of the human mind and the social world. Comte not only believed that the scientific method could be appropriated and applied to study mind and social life but that it was necessary for the development of societies. Positivism was not only a radical idea but paradigm shifting, as it had been widely held that the only thing that could be studied scientifically was the ‘natural’ world. This paradigm shift was taken up by major philosophers and social theorist from John Stuart

Mills to Emile Durkheim, facilitating an entire blossoming of social science research such as in economics, sociology, and psychology. This appropriation of the scientific method also led to the acceptance in the social sciences of natural science tenets and assumptions:

- Empirical observation is the only rational means of exhibiting the laws of the human mind.
- Universals do exist for the human organism and the social world.
- Universal truths can be established through the rational application of the scientific method.

In addition to each of these tenets and assumptions, Comte (1988) privileged the mathematical sciences as the key analytic tool for achieving precision and certainty in the pursuit of 'truth.' As Comte states, "…mathematical science is of much less importance for the knowledge in which it consists… than for constituting the most powerful instrument that the human mind can employ in investigating the laws of natural phenomena" (p. 66). As he went on to say, "It is, therefore, mathematical science that must constitute the true starting point of all rational scientific education, whether general or special" (p. 67). Hence, there emerged the quantitative imperative in the social and behavioral sciences (Michell, 1999). The call for quantification led to the later development of both statistics and measurement. Measurement became an essential goal in the study of mental and social life. In the 20th century, the positivist tenets were incorporated into the logical positivist and later post-positivist movements in the philosophy of science. The scientific method and quantification were established as central features of the guiding paradigm in contemporary social and behavioral science.

As Kuhn (1962) put forward, a scientific paradigm constitutes a set of assumptions and activities that attain sufficient popularity so as to displace all potential alternatives. In these terms, the positivist paradigm has become normal science, an almost taken-for-granted epistemological lens for knowledge production in the social sciences. Alternative assumptions and practices have been virtually buried in the mountains of research studies, textbooks, and the programmatic and disciplinary research training. This disciplinary training has given rise to what Bourdieu (1988/1984) referred to as a *homo academicus*, a species sharing a belief or faith in

objective 'truth' via the analytic tool of quantification. In effect, the positivist paradigm has emerged as a monolithic ideology, with measurement a chief hallmark[1].

History & Philosophy of Science of Measurement

The cultural and 'scientific' practice and pursuits of measurement go back centuries. Some of the more well-known earlier practices of measurement were conducted in the physical sciences, particularly physics. The practices of the physical sciences in the early 1800s, became appropriated by the positivist movement of the social sciences. Early work in anthropometry and later psychophysics became the fledgling endeavors toward what is now known as measurement and psychometric theory.

In the social sciences, measurement is often characterized as the quantification of the qualitative observation of both observable and unobservable social, cultural, and psychological constructs, processes, and objects. Stevens' (1946) classic definition of measurement as the assignment of numerals to objects or events according to rule has become the dominant model. However, as Michell (1999) reminds us, Stevens' (1946) definition of measurement is a radical departure from the classical perspective of measurement. The classical perspective defines measurement as the estimation of the ratio of some magnitude of a quantitative attribute to a specified unit of that quantitative attribute. This ratio is expressed as a real number and provides the possibility of empirically exploring the associated referent. The main distinction lies in the assignment of a numeral versus the discovery of a numerical fact. For instance, the height of a person is not the assignment of a numeral based on a rule but rather a numerical fact based on a given metric. Although Stevens' (1946) definition departed from the classical perspective of measurement, psychology and the social sciences have wholeheartedly adopted and embraced it since.

The non-critical appropriation of Stevens' operationalism in psychology (and the social sciences) was in part due to at least three forces and assumptions. These related forces and assumptions included (1) psychology's interest in being legitimated as a (natural) science, (2) the

[1] In no way are we suggesting that the measurement community is ideologically monolithic but rather that the positivist paradigm of the social sciences and policy world necessitated measurement for its endeavors of truth and objectivity. As will be discussed later there are very clear departures from positivism in the measurement community where we see much more promise.

quantitative imperative in order to be a science, and (3) the Pythagorean idea that all constructs are essentially quantitative and, as such, measureable (Michell, 1999). While the latter assumption, in particular, has been taken-for-granted in psychology and the social sciences, Michell (1999) reminds us that the question of whether unobserved mental constructs are quantifiable is still up for interrogation.

While we are sympathetic to Michell's criticisms of the taken-for-granted appropriation of Stevens' operationalism in educational and psychological measurement, we not only part with Michell's assumption of an essentialized 'natural' over a cultural but we also put into question the possibility of measuring that which is not observable. We shall later elaborate on these departures while raising further critical questions regarding the sciences of measurement. However, in order to critically question the epistemological underpinnings of contemporary measurement, more must be said about the history and philosophy of science of measurement.

Anthropometry & Psychophysics

Some of the earliest practices of measurement in the social sciences were in anthropometry and psychophysics, particularly of the measurement of mental constructs (Gould, 1996). The most notable mental construct was that of intelligence. Anthropometric methods such as craniometry were used to investigate the relationship between cranium size and human intelligence. Such anthropometric methods were employed in the research of Charles Darwin and his cousin Francis Galton, who later became known as the father of both psychometrics and the eugenics movement. Although the later work of cultural anthropologist Franz Boas discredited the eugenic aims of anthropometry, the appropriation and employment of methods from psychophysics continued.

Psychophysical methods were also used to study the relationship between stimuli and mental responses. Gustav Fechner's psychophysical studies - employing the measurement of mental process - were among the earliest works in psychology. Fechner's psychophysics later influenced one of the major thinkers in psychometrics, L.L. Thurstone, whose approach to measurement was based on the idea of comparative judgment. While the methods of psychophysics continued to develop in such areas as mathematical psychology and sensation and

perception research, they also led to the development of classical test theory and the search for truth in the quantification of unobservables.

True Scores & Classical Test Theory

One of the most widely employed theories of measurement in psychology and the social sciences is classical test theory. First developed in 1888, its basic model is Observed Scores = True Scores + Error. The error term is for measurement error and, in part, is what makes this a measurement model. It is a very straightforward model. As a population model, it also suggests that the observed scores in a random sample of the population is a function of both true scores of the construct, with some degree of random measurement error over the population. The measurement error is assumed to be random and non-systematic over the population and is a population specific estimate. Thus, it should not be interpreted or generalized beyond the population in which it was estimated on. The model specifies no relationship between the true score or unobserved construct, and the observed behavior. And, moreover, the model only accounts for person parameters, and does not account for item parameters. In other words, the observed and true scores are parameters estimated for the person (e.g., estimated ability) and the test (e.g., error), but there are no parameters estimated for the items themselves.

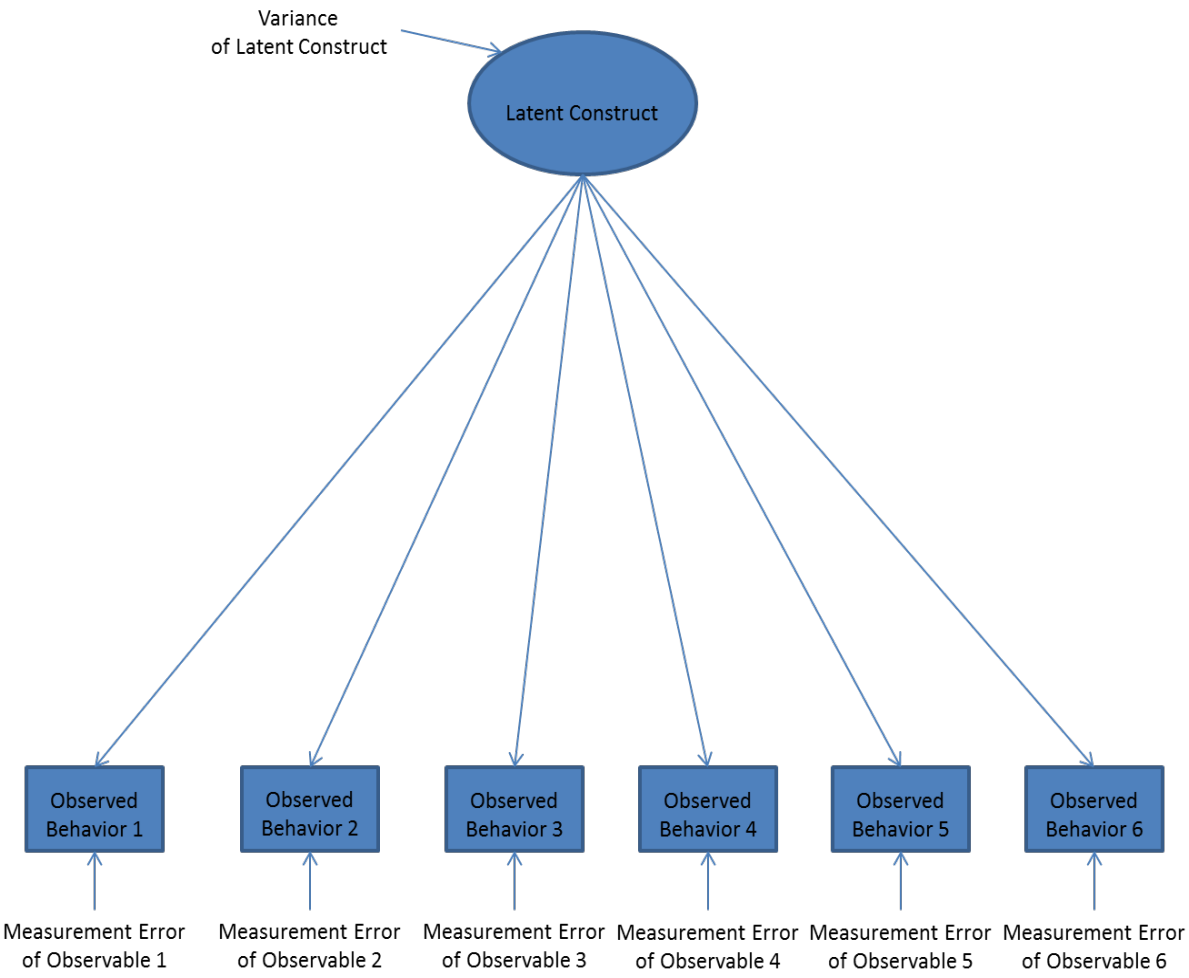
Most important, the true score assumes that there is a true construct with a certain degree of random error in the population. The positing of true scores makes very strong metaphysical assumptions, as the true score is essentially based on an unobserved construct. The often used realist interpretation of true scores is, as Borsboom (2005) states, "a metaphysical entity of the worst kind." In fact, an operationalist lens to the true score, which holds that the theory of the true score is synonymous with the operations that have been measured, would lead to a more reasonable interpretation of the true score. At the same time, it would encounter other limitations such as no two instruments can measure the same construct. Given its limitations as being a population model, its non-specification of the relationship between the true scores and observed behaviors, and its inability to account for both item and person parameters in the model, latent variable models were subsequently developed.

Latent Variables and Latent Variable Modeling

In the early 1900s, Charles Spearman developed what became known as factor analysis. As a method that rests on correlations, factor analysis models the correlation structure of three or more observed/manifest behaviors, relying on the assumption that what underlies this covariance structure of behavior is a latent construct or variable. Latent variable modeling further developed into confirmatory factor analytic models, item response theory, the Rasch model, latent class models, and latent profile analysis, among others. Latent variable models are among the most widely developed measurement tools in psychometric theory. In contrast to the classical test theory model, the latent variable model accommodates both person and item parameters, specifies a relationship between the latent variable and the observed behavior, and is not limited to a population model. Thus, the models allow for flexibility in inferences to be made about each item, the test, and the estimation of the latent variable. Measurement error is specified for each item, as well as the test, and can be estimated for subgroups with the conditional standard error of measurement.

As observed in figure 1, the latent construct is assumed to be associated with and materialize in the observed/manifest responses. In fact, based on the estimated model the conditional probability of an observed/manifest response is a function of the test-taker's theta (i.e., latent score or ability). Latent variable models also evaluate the fit of the model to the data. Thus, a hypothesized model is specified, data is collected, then the hypothesized model is evaluated based on its fit to the empirical data. The earlier latent variable models were evaluated on the basis of the traditional positivist framework for hypothesis testing, but later models (e.g., structural equation models or item response models) adopted the Popperian post-positivist lens of falsifiability. Under either lens of hypothesis testing, a realist interpretation of mental events is required in order to assume that there is a 'true' model and, as such, a 'true' construct.

Figure 1. Classic Factor Analytic Model



Given that the operationalist view would not hold for many of the procedures of latent variable modeling (e.g., computer adaptive testing where each test-taker is administered a different set of items in order to measure the same latent construct), a mental realist perspective is often presumed in the interpretation of the latent variables. That is, the latent variable is supposed to refer to a real mental entity. In addition, given that the interpretation to measurement error for latent variables is analogous to true scores – where measurement error is assumed to be random error of the 'true' construct – there is a trace of the assumption of a fixed and stable unobserved construct.

This (post-) positivist understanding of the relation of the latent variable to the observables as well as realist interpretation of the latent variable are also profoundly haunted by metaphysical assumptions. As an assumption, the metaphysics privileges the presence of the observable over the absent presence of the unobservable mental processes, assuming that the present observable fully and purely signifies the meaning, intentionality, or consciousness of the unobserved internal processes, or mental constructs. Thus, the presence of the observable is assumed to be pure whereas the absent presence of the unobserved mental processes is assumed to be impure and imperfectly signified by the mental construct. What this assumption overlooks is the residue, or the trace of prior experience in the present observable rendering it also contaminated and impure. That is to say, the present observable is always signifying something beyond itself, not quite specified. As such, the exact referent of the observed behavior is always 'to come'. Psychological measurement is essentially an act of transferring meaning regarding mental processes onto observed processes and behaviors in the social world. Even when we think we have captured the meaning of those mental processes they are always ambiguous, shifting, and slipping away. There is no possibility of accessing meaning, intentionality, and consciousness of inner mental processes via the signifiers of observable behaviors.

The above deconstruction of the metaphysics of latent variables has direct implications to validity. As a central topic of measurement, scholars such as Cronbach and Meehl (1955), Messick (1989, 1994, 1998), Kane (1992), Mislevy (2008), and Moss (1994) among others have attempted to address different critical questions pertaining to the concept of validity. Messick (1994) argued that validity is a value judgment that consists of six unifying aspects: content, substantive, structural, generalizability, external, and consequential. Following Messick's contributions, others (Kane, 1992; Mislevy, 2008) have pushed for a model-based and argument-based reasoning approach to validity where the models are understood to be imperfect. The above deconstruction suggests that validity is an impossibility with unobserved constructs. While we are sympathetic to Messick's concern for the social consequences of score interpretation and use (i.e., consequential validity) we are suspicious of any measurement endeavors that claim to have any privileged access to unobserved cognitive or psychological processes. In any case, content and external validity, because they rest on ungrounded interpretation, are an impossibility.

Measurement Scales & Representational Models

The third measurement model is based on S.S. Stevens' (1946) measurement scales known as representational models. Stevens' measurement scales (i.e., nominal, ordinal, interval, and ratio) widely influenced the social sciences, particularly with respect to measurement and statistical analysis. These measurement scales were also employed by the logical empiricists of the Vienna Circle in order to develop *representational models* of measurement. Influenced by psychophysics and developed to address the principles of fundamental measurement as specified in physics, representational models are based on a theory of the relationship between objects. Thus, the observed behavior is not measured but rather the 'noticeable' difference between objects is assessed. The 'noticeable' difference implies human perception and thus necessitates interpretation. Representational models are not measuring some unobserved construct but rather only observed differences. Representational models are better able to conform to the classical perspective of measurement in exploring the ratio of intensities or magnitudes of perceived differences. These 'noticeable' differences are then quantified into measurement scales and constituted within a mathematical model, as opposed to a statistical or probability model as in latent variable models.

These models are most employed in mathematical psychology and psychophysics and are prominent in the work by Khaneman and Tversky (1974) on judgement and decision making theory, as well as in research on sensation and perception. In addition, the Rasch model (1960/1980), which is a special case of item response theory, has been argued by scholars such as Ben Wright (1999) to have been based on the principles of fundamental measurement theory. Ben Wright's argument is based on the Rasch models employment of Luce and Tukey's (1964) theory of conjoint measurement. However, because the Rasch model does account for measurement error and estimates *theta*, or a 'true' construct, it makes assumptions of a 'natural' or 'real' object of measurement. In effect, the Rasch model also often falls trap to the realist metaphysics of latent variables.

There are several limitations with representational models. Various factors, affordances, and influences can affect subject perception and interpretation 'in-situ' and, as such, need to be examined in a 'controlled' environment. Thus, these measurement models are often established within experimental conditions (with the exception of the Rasch model). In addition, because

perceived differences conform to the classical perspective on measuring psychological attributes, they still fall trap to interpretations – in this case the unobserved process of human perception. Again we have metaphysical speculation. This is analogous to the procedures of an eye examination, with the subject's reports susceptible to influences of motivation, attention, concern with social appearance, and so on. Moreover, the exploration of perceived differences says nothing about a measured person's creative skills, quantitative skills, literacy, knowledge of science, history, etc. It is on these latter grounds that the representational models of measurement fall short of the goals of educational assessment.

Emerging Approaches to Measurement

The past few decades of development in theories of learning and cognition have brought about new, postmodern perspectives in understanding. These include socio-cultural and situative perspectives on learning, knowledge making, and human development. These perspectives have been informed, for example, by the works of Lev Vygotsky, cultural psychology, and educational developments in sociology, anthropology, and linguistics. The work of scholars such as Robert Mislevy, Pamela Moss, Edward Haertel, and others (see Moss et. al., 2008) have also led to new approaches to measurement. In particular, Robert Mislevy has developed what he has referred to as *a constructivist-realist, neopragmatic* postmodern theory of measurement. In his essay titled "Postmodern Test Theory", Mislevy (1997) leans on neo-pragmatic, critical legal studies to advance a view of measurement that maintains the analytic utility of measurement modeling. In this approach, he postulates two essential elements:

(1) Understanding the important concepts and relationships in the learning area in order to know the aspects of students we need to talk about in the universe of discourse our assessment will generate and (2) determining what one needs to observe and how it depends on students' understandings, so as to structure assessment settings and tasks that will provide evidence about the above mentioned aspects of students' understandings. (Mislevy, 1997, p. 190)

Thus, not only does this approach emphasize developing an understanding of the repertoire of information we need to know about a student in a particular area of learning but also situating this in the students' localized knowledge and understanding. This more situated approach is better able to account for the affordances of the learning environment, as well as

broad variations in socio-cultural particularities. Thus, the statistician would not treat probability models of measurement as having a fixed meaning "but as provisional understandings with situated meanings, based on previous experience" (Mislevy, Moss, & Gee, 2009, p. 89). Additionally, Mislevy attempts to avoid the metaphysical construction of *theta* (the measured construct) by situating theta not inside the head of the test-taker but the tester (personal communication on February 28th, 2012). By situating theta inside of the head of the tester, item response theory models move away from their misleading mental realism. Mislevy also argues for the use of measurement models as a tool for evaluative discussion and not as a means of determining the truth about their objects. As such, the discursive tools of measurement should be triangulated with various qualitative approaches to deliberate on what the student or test-taker 'knows' or is able to do.

Others have also similarly followed in this liberating line of work. For example, Valerie Shute's (Shute & Spector, 2008) work on stealth assessments has led measurement toward new possibilities. Stealth assessments are seamlessly woven into the fabric of a learning environment (e.g., virtual world) so as to be unnoticed or unobtrusive to the learner, while employing machine-based, non-psychological inferential tools (e.g., estimating values of evidence-based competencies across a network of skills). These new possibilities have attractive features, even if harboring limitations. For example, while Mislevy's re-situating of theta as "inside the head of the tester" avoids a metaphysics of the object, it simultaneously re-positions the site of metatheoretical conjectures, and lends itself to un-reflective, top-down postures toward what is measured. The shifting of the theta continues to sustain the measurement gaze on mental or cognitive processes. The latter point is critically important given that Mislevy is arguing for the utility of a 'modeling' based approach, an approach that inherently normalizes knowledge at the cost of legitimating particular forms of knowledge over and against others. Moreover, modeling based approaches still rest on the act of freezing the rapid movements and fluidities of the cultural world, more summative in their effects as opposed to formative. Despite these remaining quandaries Mislevy's neo-pragmatic post-positivist approach does point toward promising possibilities.

Concluding Reflection: Measurement and Cultural Meaning

The preceding journey through the history and philosophy of science of measurement reveals a number of epistemological quandaries. We have focused especially on the way in which the predominant orientation to testing (particularly in the policy world) misleadingly presumes a realist stance toward "measuring the mind." Further, even when there is no presumption of a mental world behind the score, one cannot escape problems of value biased readings of measurement outcomes. Simply in the labeling of a performance measure, for example, there are ideological and political loadings. Is a “test of reading ability” simply this, or is it a test of home environment, economic privilege, or parental influence? While many in the measurement community have moved from the realist toward more of a constructivist-realist perspective this has been slow to take hold in the policy community. Related to and beyond some of the quandaries are other critical questions concerning the construction of knowledge through measurement. For one, measurement sciences sustain the assumptions of universality at the cost of the multiplicity in social and cultural particularities. Modernist approaches to measurement – which continue to be the dominant approach in policy and practice – make strong assumptions about the minds and actions of those to be measured. The act of defining a measured construct is a constituting and fixing of these minds and actions, privileging the authority’s judgments about those who are measured. This defining and constituting creates a cultural universe of meaning that simultaneously excludes and de-legitimates alternatives.

With the exception of cognitive diagnostic or mixture models, the universal/general aims of measurement modeling have also been at odds with the social-relational process of education. Education is essentially a social process (Dixon-Román & Gordon, 2012; Varenne, 2009), and as such, is in-situ and reflects the needs, values, and expectations of local populations. These processes differ dramatically across the nation, according to diversity in ethnicity, education, religion, urbanity, political investments, and so on. As Gee puts it, “Assessing people who are not in the same “discourse community” … is meaningless and sometimes unethical” (Mislevy, Gee, & Moss, 2009). This is what Freedle (2003) argued with the cultural familiarity hypothesis regarding group differences in performance on the verbal items of the SAT. This hypothesis suggests that the more localized and situated understandings of language are often in conflict with and incongruent to the dominant cultural understanding reflected in standardized

assessments such as the SAT. Mislevy argues for the utility of measurement models that are situated in discourse communities and interpreted as provisional and not fixed (Mislevy, 1997; Mislevy, Gee, & Moss, 2009). Necessitated here is dialogue with and among participants in the local communities. As we shall advance in the second paper, this is an attractive approach to working with the tensions between the universal aims of measurement and the particularities of local social and relational processes.

We are also concerned with the ideological investments that inhabit the production and reading of measurement instruments. Although empirical science has laid claim to value neutrality, insisting that facts are value free, recent decades have illuminated the misleading character of such claims. What is counted as "fact" is always located within a community tradition or cultural history, and reflects its particular values (Kirby, 2011; Poovey, 1998; Rotman, 2000). This is also to point out that measurement too is the product of socio-cultural process. As mathematician and philosopher, Brian Rotman (2000) has described, mathematics is no less a language than the spoken and written language of the culture. It is a cultural practice, and its meanings are socially produced and contestable.

We must ask, then, whose values are inherent in the hierarchies of performance embedded in the measurement of student behavior, and why should they be privileged over the many other values that circulate within the society? The ideological saturation of measurement does not render measurement useless; it simply demands that we deliberate more broadly and inclusively about its implications for education and society.

This is also to note that although measurement can be employed as a form of assessment, measurement is only a limited special case within the larger world and practice of assessment (Shepard, 2000). Such questioning also sets the stage for considering alternative epistemological possibilities. It is thus that we move in the next paper, to consider a social constructionist epistemology as a promising alternative to the empiricist tradition of measurement.

References

Berman, M. (1988). *All that is solid melts into air: The experience of modernity*. New York NY: Penguin Books.

Bourdieu, P. (1988/1984). *Homo academicus*. Stanford CA: Stanford University Press.

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. New York NY: Cambridge University Press.

Comte, A. (1988). *Introduction to Positive Philosophy* [edited by Frederick Ferré]. Indianapolis IN: Hackett Publishing Company, Inc.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Derrida, J. (1972/1982). Signature event context. in J. Derrida *Margins of Philosophy*. Chicago IL: University of Chicago Press.

Dixon-Román, E., & Gordon, E.W. (2012). *Thinking Comprehensively About Education: Spaces of Educative Possibility and their Implications for Public Policy*. New York NY: Routledge.

Eitzen, D.S., & Zinn, M.B. (2011). *Globalization: The Transformation of Social Worlds*. Beverly, MA: Wadsworth Publishing.

Gould, S. J. (1996). *The mismeasure of man* (Rev. ed.). New York: Norton & Company Inc.

Hardison, O.B Jr. (1995) *Disappearing through the skylight: Culture and technology in the twentieth century*. New York: Penguin.

Kane, M. T. (1992). An Argument-Based Approach to Validity. *Psychological Bulletin*, 112(3), 527-535.

Kirby, V. (2011). *Quantum Anthropologies: Life at large.* Durham, NC: Duke University Press.

Khaneman, D., & Tversky, A. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124-1131.

Kuhn, T. (1962/1996). *The structure of scientific revolutions* (3rd Ed.). Chicago, IL: University of Chicago Press.

Luce, R.D. & Tukey, J.W. (1964). Simultaneous conjoint measurement: a new scale type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1–27.

Maffesoli, M. (1996). *The Time of the Tribes: The decline of individualism in mass society*. Thousand Oaks, CA: Sage Publications.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, 45(1/3), 35-44.

Michell, J. (1999). *Measurement in Psychology: A Critical History of a Methodological Concept.* New York NY: Cambridge University Press.

Mislevy, R. (1997). Postmodern test theory. In A. Lesgold, M.J. Feuer, & A.M. Black (eds.) *Transitions in work and learning: Implications for assessment.* National Academy Press.

Mislevy, R.J. (2009). Validity from the perspective of model-based reasoning. In R.L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications.* Charlotte, NC: Information Age Publishing.

Mislevy, R., Gee, J., & Moss, P. (2009). On qualitative and quantitative reasoning in validity. *Generalizing from educational research*. New York NY: Routledge.

Moss, P. A. (1994). Can There Be Validity without Reliability? *Educational Researcher*, 23(2), 5-12.

Moss, P.A., Pullin, D.C., Gee, J.P., Haertel, E.H., & Young, L.J. (2008). *Assessment, Equity, and Opportunity to Learn*. New York NY: Cambridge University Press.

Poovey, M. (1998). *A History of the Modern Fact*. Chicago IL: University of Chicago Press.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago IL: The University of Chicago Press.

Ritzer, G. (2009). *Globalization: A Basic Text*. Walden, MA: Blackwell Publishing.

Rodgers, D.T. (2011). *Age of Fracture*. Cambridge MA: Belknap Press of Harvard University Press.

Rotman, B. (2000). *Mathematics as sign: Writing, imagining, counting*. Stanford, CA: Stanford University Press.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher* 29(7):4-14.

Shute, V. J., & Spector, J. M. (2008). "SCORM 2.0 White Paper: Stealth Assessment in Virtual Worlds." Retrieved May 16, 2011, from http://www.adlnet.gov/Technologies/Evaluation/Library/AdditionalResources/LETSIWhitePapers/Shute–StealthAssessmentinVirtualWorlds.pdf

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 667-80.

Varenne, H. (2009). Educating ourselves about education—Comprehensively. In H. Varenne, E. W. Gordon, & L. Lin (Eds.), *Perspectives on Comprehensive Education Series: Vol. 2. Theoretical perspectives on comprehensive education: The way forward* (pp. 1–24). Lewiston, NY: The Edwin Mellen Press.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know* (pp. 65-104. Hillsdale, New Jersey: Lawrence Erlbaum Associates.