The Gordon Commission
on the Future of Assessment in K–12 Education

Gordon Commission

Fellow Synthesis Paper

Amanda Walker Johnson

Assistant Professor
University of Massachusetts Amherst

> But one of the things that has to be faced is, in the process of wanting to change that system, how much have we got to do to find out who we are, where we have come from and where we are going. . . . I am saying as you must say, too, that in order to see where we are going we not only must *remember* where we have been, but we must *understand* where we have been.
>
> - Ella Baker, (quoted in Moses and Cobb 2001: 3)

This paper reviews and synthesizes the following 10 papers commissioned by the Gordon Commission on the Future of Assessment:

1) Ana Marie Cauce and Edmund W. Gordon, *Toward the Measurement of Human Agency and the Disposition to Express It*
2) A.S. Meroe, *Democracy, Meritocracy and the Uses of Education*
3) Rodolfo Mendoza-Denton, *A social psychological perspective on the achievement gap in standardized test performance between White and minority students: Implications for assessment*
4) Hervé Varenne, *Education: Constraints and Possibilities in Imagining New Ways to Assess Rights, Duties and Privileges*
5) Ezekiel Dixon-Roman and Kenneth Gergen, *Epistemology and Measurement: Paradigms and Practices I. A Critical Perspective on the Sciences of Measurement*
6) Kenneth Gergen and Ezekiel Dixon-Roman, *Epistemology and Measurement: Paradigms and Practices II Social Epistemology and the Pragmatics of Assessment*
7) Andrew Ho, *Variety and Drift in the Functions and Purposes of Assessment in Education*
8) Robert Mislevy, *Four Metaphors We Need to Understand Assessment*
9) Carl Kaestle, *Testing Policy in the United States: A Historical Perspective*
10) Clifford Hill, *Assessment in the Service of Teaching and Learning*

Cauce, Ana Marie and Gordon, Edmund W. (2012). *Toward the Measurement of Human Agency and the Disposition to Express It.*

Findings

Given evidence of a strong link between preconditions for agency and academic achievement, this article considers what conditions are necessary for the expression of agency and how such conditions can be assessed. Synthesizing the work on agency from social-psychological and economic perspectives, the authors define human agency as “the capacity and disposition to recognize and act in one’s own best interest and that of chosen others” (p.6). Additionally, they arrive at several propositions about agency:

- Human agency is multidimensional, measurable both in general and in specific domains.

- There exist facilitative conditions for agency, or *capacity*, the ability to make free choices, access resources, and exercise autonomy. A key question is how to measure context without either erasing variability or generalizing exceptional cases.
- Measures of agency cannot be equated with measures of achievement or outcome, but must be considered as measuring "capability-in-action" (p.14). An example is the use of portfolio assessments, such as scholarship competitions.
- While agency does not always find positive moral expression, the authors limit their definition of agency to "acting in the service of the greater good" (p.17).

Recommendations

- Emphasize to youth the links between effort, goals, and fulfillment.
- Redefine the purpose and role of education as producing and developing agency.
- Counter the ways that testing can "crowd out self-reflection, forethought, and creativity."
- Develop holistic assessments of agency, such as portfolios.
- Consider the assessment of collective agency.

Commentary

Varenne's interaction theory merges with this piece in that, as systems crowd out agency, people will seek other ways of expressing their agency. In Mendoza-Denton's work, while toxic environments can be seen as limiting agency, an incrementalist (versus fixed/entity) view of intelligence can be seen as agentic. Levels of self-efficacy can be viewed vis-à-vis levels of protection of social identity and self-esteem, and the conditions that trigger self-protection are exactly those that limit capacity. Ho's distinction between assessments of the learner and of learning, and the teacher and teaching provides a framework for understanding intellective competencies, as well as consideration of the conditions for agency for teachers as well as students. Hill's meaning of "authentic" in speaking of digital assessment implies *agentic*, and leads to a consideration of how digital archives (process logs and chains of feedback) might provide a source for assessing agency. Meroe's work suggests a need to contextualize the ways in which schools and testing crowd out agency within a broader history of meritocracy. Considering Dixon-Roman and Gergen, what epistemologies support conditions of limiting or enabling agentic expression? Perhaps, the dichotomization between achievement and self-realization/creativity, noted by Cauce and Gordon, is a product of historical processes and structures of power.

Meroe, A.S. (2012). *Democracy, Meritocracy and the Uses of Education.*

Findings

In the United States, conceptions of democracy and meritocracy tend to be conflated, in large part due to their common critique of aristocracy and birth privilege, as well as common conceptions of individual autonomy. Nevertheless, there exist "shadows" of both concepts that further social inequality. Both democracy and meritocracy were developed in the context of several tensions, between: 1) full participation and notions of a deserving elite; 2) majority-rule on the one hand, and individual and minority rights (and protections) on the other hand; and 3) expansions of freedoms (in the United States) in the context of the brutalities of slavery, colonization, and gender subordination. The assumptions of the American Dream (autonomy, the Protestant work ethic, and equal opportunity) supporting the conflation of democracy and meritocracy have coexisted with the explanation of social inequalities through ideologies of Social Darwinism and market-based logics (or the equating of *freedom* with the free market).

The resilience of the shadows of meritocracy, despite well-known, persistent inequalities can be explained by the ability of meritocracy to be a moral cover for inequality, but also the alignment of meritocratic perspectives with liberal capitalism. Both the acceptability of the injustices of capitalism (i.e., "that's life") and the tokenization of successes of underprivileged groups also support meritocracy. Though social inequalities and unequal distribution of resources preclude the "completeness" of meritocracy, contemporary neoliberal ideologies oppose governmental intervention to counter material inequalities. These work against collectivist-democratic perspectives for policymaking.

Recommendations

- Employ "critical recognition" as a methodological tool, which includes 1) historiography of concepts (social epistemology), 2) accounting for the "shadows" and historical consequences, and 3) mapping the distribution of resources and recognition.
- Take responsibility for promoting education for all as a social good.
- Counter individualism with collective democracy and (re)distributive justice.
- Look to other countries, such as Finland, for examples of how to balance assessment, governmental support, and a collective ethos for educational systems.

Commentary

One key role that Meroe's piece can play for assessments is to provide a framework for evaluating assessments: what are the shadows, underlying tensions, material conditions, and aspects (and distributions) of recognition embedded in assessment practices. I think this answers the challenge by Ho to develop a methodology for anticipating "purpose drift," as well as charting the "structural incentives and historical precedents" for purpose drift.

In terms of epistemology, Mislevy contends that the "bottom line" is that paradigms can be rearticulated, shifted from their genealogical precedents or origins (p.27) — i.e., measurement doesn't always have to be positivist, or meritocracy doesn't always have to be an arm of liberal (or neoliberal) capitalism. To this point, one question that Mislevy and Kaestle might pose with regard to Meroe's piece is whether or not there are (everyday) conceptions of meritocracy that have challenged the reproduction of privilege, Social Darwinism, and market-logics. Consider, for example, the case of W.E.B. DuBois and his role in the institutional development of African American college graduates and scholars, as well as Black scholarship. While DuBois's conceptions of the Talented Tenth have been the subject of various critiques, in any case, his conceptions of meritocracy were rooted in the collective ethos championed by Meroe. Following this line of thought, Varenne's paper suggests asking how critical, everyday rearticulations of meritocracy can be leveraged to develop more equitable educational experiences and assessments.

Mendoza-Denton, Rodolfo (2012). *A social psychological perspective on the achievement gap in standardized test performance between White and minority students: Implications for assessment.*

Findings

The achievement gap can be explained by "toxic environments." Social psychological studies show that stigmatization can ignite the development of self-protective measures: disidentification and "attributional ambiguity," whereby social rejection can be attributed to discrimination/racism/sexism, rather than to the self. Mediating processes of attribution and identification are a sense of individual responsibility and the availability of social narratives to contextualize discrimination. Subtle racism can trigger attributional ambiguity, and perceptions of discrimination can lead to rejection of positive feedback. However, if students trust that race is

not being held against them, then they tend to attach self-concept to feedback. Protective measures for self-esteem may not protect against anxiety, lack of control, and impacts on physical health.

"Stereotype threat" has been reconceived as a threat to one's social identity. The latter threat can transform disidentification with a specific situation to disidentification with an entire domain. Research shows that being made aware of the racial, gendered, caste-based, and socioeconomic stereotypes is linked to underperformance, while those with low sensitivity to race (rejection) tend to have higher academic achievement. Experiences of discrimination, social rejection, devaluing of identities, and lack of belonging can increase sensitivity to race or other status, and manifest themselves in lowered achievement and academic disengagement. There are two conceptions of intelligence: 1) entity — fixed, dispositional; and 2) incremental — the sense that intelligence is malleable and can "grow." Self-preserving strategies that "entity theorists" adopt mirror those of minority students under stereotype/social identity threat.

Recommendations

- Reframe from the question of biased or unbiased testing to one that considers threatening or nonthreatening environments.
- Interventions should target these areas: shift to incrementalist view of intelligence; valuing of social identities; and creating a sense of belonging.
- The testing industry also could promote the incrementalist view, reconceptualize instruments without *a priori* group differences, advocate for the end of tracking, and relationally diversify itself as a field.

Commentary

Like the reports by Cauce and Gordon, Meroe, and Gergen and Dixon-Roman, this paper, perhaps most of all, draws our attention to the psychological and even biological well-being of students. It also points to what might be called "higher order" kinds of unfreedoms, as well as the potential for incrementalist views of intelligence to empower students. A view of mechanisms of self-protection and social identity threat may help explain what Gergen and Dixon-Roman describe as the results of high-stakes testing environments.

Varenne, Hervé (2012). *Education: Constraints and Possibilities in Imagining New Ways to Assess Rights, Duties and Privileges.*

Findings

The institution of the public school has failed to bring about social equity. While school has increased access, it has failed to close the gap or to eliminate birth privilege, instead reproducing it. The school can function as a barrier to, rather than a purveyor of, justice and opportunity. In fact, the current state institution of the school may be in crisis, hindering progress in the next century and becoming too narrowly vocational, or tailored to meet the needs of industries. One explanation for the failures of the school is that it has not taken into account "interaction theory" and has not recognized *agency* and the educative quality of everyday life moments.

Methods of self-education exist, as well as systems of ethno-education, ethno-assessments, curriculum, and ethno-pedagogy. Concrete examples of these systems have been explored; in particular, analyses of blog responses, multiple player games, and parents' advocacy for and education of autistic children. These analyses explore chains of responses, everyday assessments of competence and linguistic abilities, and personal investments in curriculum, pedagogy, and assessment.

Recommendations

- Challenge the dichotomization of the "educated" and "uneducated" that ultimately imposes deficit-models upon those with lower levels of formal education.
- Use an analysis of ethno-assessments to imagine alternative routes to the granting of privilege.
- Leverage what we know about everyday assessments to reorganize school-based assessments, considering questions such as who, what, when, and to what effect, as well as the "life-changing" stakes of assessments (p.22).
- The state should protect the right to education and not devolve its responsibility to provide free and public education to all. However, the state should not use school-based assessments and certification (or degrees) as a means to grant career privileges. While it is reasonable to ask for assessments for selection and granting privilege, the school does not have to be the site for those assessments.

Commentary

One common theme between Varenne's and Meroe's paper is the goal of the elimination of birth privilege. For Meroe, this goal lies at the core of the acceptance (and conflation) of meritocracy and democracy. For Varenne, this is a measure for the efficacy of schooling. A dialogue between Meroe's and Varenne's reports might consider how the state, as a bloc of multiple and often contradictory alliances, mediates tensions extant in "the Commons": between groups who ally to challenge privilege and those who ally to protect that privilege; e.g., the case of California tax revolts (Kozol, 1991). One question is how Varenne conceives of job placement tests, required to obtain employment in public and private sectors, which function in large part outside of the school.

Varenne's discussions of social interaction align with Cauce and Gordon's discussion of agency and Mislevy's sociocognitive view of *practice*. Mislevy's report leads to questions of how to evaluate and/or improve everyday assessments according to the kinds of practices, evidentiary arguments, feedback loops, and measurements they produce. Additionally, how might these everyday assessments be validated considering Ho's framework (measurements and influencing)? Further, what can regulate or govern the possible "purpose drift" of everyday assessments?

Like Dixon-Roman and Gergen, who challenge the philosophical foundations of assessments, Varenne critiques a fundamental tenet of U.S. ideologies of schooling: that of the state control over the granting of privilege. The critiques of neoliberalism by Meroe and Gergen and Dixon-Roman suggest questioning how Varenne's can be distinguished from neoliberal critiques of the State as holding a "monopoly" over education.

Dixon-Roman, Ezekiel and Gergen, Kenneth (2012). *Epistemology and Measurement: Paradigms and Practices I. A Critical Perspective on the Sciences of Measurement.*

Findings

In the context of "seismic" shifts in communication due to globalization, traditional testing practices tend to "freeze" the social order instead of acknowledging: 1) the multiplicity and particularity of knowledge and learning practices, 2) shifts in demands for knowledge domains, and 3) shifts in pedagogical practice. This tendency to freeze is due in part to the

paradigmatic assumptions upon which the current, traditional testing practice is based, namely the positivist paradigm.

The positivist paradigm is based on the application of measurement and physics to the study of human intelligence and social life, following Comte in the 19th Century, that emphasized empiricism, the establishment of universal truths, rationality and the scientific method, and the privileging of quantification. The authors find that the definition of measurement by Steven (1946), so influential to the development of educational measurement, constituted a "radical break" from classical definitions of measurement: shifting from measurement as mathematical derivation (or "discovery of numerical fact") to measurement as enumeration (or assigning numerals to objects) (p.5). The three assumptions that allowed for this redefinition to become hegemonic in psychology include: 1) the search for disciplinary legitimation as a science, 2) the quantitative imperative, and 3) Pythagorean assumptions about the measurability of all concepts.

As models for testing have emerged that intended to address the weaknesses of earlier paradigms — namely, classical test theory, latent variable modeling, representational modeling, and constructivist-realist, neopragmatic postmodern testing theory — they still share metaphysical assumptions about knowledge and the reification of mental processes; assumptions about the universality of measurements; the privileging of hierarchical, top-down judgments of capabilities; and conflicts with educational practice.

Recommendations

The authors suggest adopting a perspective respecting multiple vocality and conducting an inventory of the "ideological investments" supporting practices of assessment. They provide a second paper with an alternative constructivist paradigm of assessment.

Commentary

Like Varenne's, this piece challenges the conceptions fundamental to current testing practice. Alongside the reports of Meroe and Kaestle, it suggests that the positivist application of physics to social and mental processes has coincided with the faith in meritocracy (as a more progressive and scientific approach to granting privileges) and has supported the development of technical experts who constituted the managerial class. At the same time, the theoretical

transformations necessary to apply classical measurement to social life make up the taken-for-granted assumptions in traditional or widespread testing, shedding light on what Ho describes as reification and naming fallacies.

Gergen, Kenneth and Dixon-Roman, Ezekiel (2012). *Epistemology and Measurement: Paradigms and Practices II Social Epistemology and the Pragmatics of Assessment.*

Findings

There are three major lines of critique of, or challenges to, the empiricist paradigm constituting the *traditional* system of testing in the United States:

- ideological analyses, stemming from Marxist analyses, critiquing and (ethnographically) interrogating the value-neutrality claim of science; instead, finding the values and ideologies inherent in scientific practice
- linguistic/literary theory, stemming from Saussure and Derrida, which emphasize the arbitrariness of signs, social conventions of referents, as well as the narratives and metaphors embedded in scientific practice
- Social epistemology, particularly influenced by Kuhn, which challenges the idea that scientific knowledge has progressed toward some ultimate truth, arguing instead that the historical trajectories or evolution of scientific practice were based on shifts in paradigms, and the unseating, development, and maintenance of norms in scientific communities. Social epistemology, then, calls attention to the historical and cultural situatedness of science and "tradition[s] of understanding" that delimit understandings (p.4)

These critiques do not preclude standardized testing or science, but call for what Sandra Harding (1993: 17-19) might call "strong method and strong objectivity."

These lines of critique then call into question the ideological *investments* that support testing and that testing supports, in particular neoliberalism (the tendency to reduce social life to market logics) and individualism (both the reduction of complex and social phenomena to the individual and the ethos of selfishness). In the current environment, the effects of testing on patterns of behavior include hierarchies that support alienation and create environments of distrust, antagonism, and dishonesty; "hierarchies of worth" that produce labeling effects; and counter-democratic practices that suppress pluralism and particularity. Effects of testing on educational process include narrowing of curriculum and pedagogical methods; reducing motivation and engagement in both teachers and students; and negatively shaping parents' views

of their children. Alternatives exist that embrace a sociocultural, constructivist perspective including:

- empowerment and dialogic (or participatory) evaluation of schools, by which communities can engage in self-evaluation; honoring multivocality
- appreciative evaluation that decenters the "problem" as the starting point/impetus for reform, and instead uses what local communities value and prize to develop goals and evaluative tools
- sociocultural, situative assessment, such as Mislevy discusses

Recommendations

- Move away from the following: assessment as selection and prediction; reification of internal mental structures; establishment of hierarchies of reward and punishment; and standardization of curriculum as educational policy.
- Propose assessment that is dialogic, multivocal, constitutive of multiple criteria, formative, and professional-developmental for teachers and administrators.
- Keep the standardized testing system, but instead of mandating testing, expand the availability and kind of testing to communities, training them in participatory evaluation (as a form of accountability).

Commentary

Hill's digital assessments could be characterized as constructivist, as they seek increased authenticity, social relation, and student voice. Kaestle's analysis of the history of testing urges caution in arguing about the causality of testing, and might suggest more fully explicating which type of testing in what historical period is responsible for what effects, for example IQ testing in the 1920s, testing such as ITBS, minimum-competency tests of the 1970s, or high-stakes testing in the accountability era. Gergen and Dixon-Roman also provide a framework for considering the extent to which Mislevy's social and situative perspectives on assessment have, or can, become paradigmatic in educational measurement.

Ho, Andrew (2012). *Variety and Drift in the Functions and Purposes of Assessment in Education.*

Findings

Defining purpose (use and interpretation) is key to validity, as current methods of validation are inadequate. One solution is to consider and perhaps synthesize four frameworks by Haertel and Herman, Haertel, Pellegrino, et al. (NRC), and Kane for establishing purpose. Some of the key points in the frameworks include distinguishing between measuring and influencing purposes, as well as distinguishing levels and sites of intervention: student, teacher, schools, school systems, and the public.

Purpose drift, the "proliferation of purposes" especially in high-stakes environment, is in part made possible by the cultural acceptance of "reification …, naming …, and ecological fallacies," linked to quantification (p.13). While formative assessments should be pursued, summative assessments often problematically become conflated with high-stakes testing.

Recommendations

- Validity frameworks need to pay more attention to differentiating assessment of learners and learning, as well as differentiating that of teachers and teaching.
- It is important and necessary to clearly define the targets for validation.
- We need to "[understand] the structural incentives and historical precedents" for purpose drift, as well as develop the means for anticipating and stemming purpose drifts (p.14).
- Ultimately, we need to be more proactive in challenging the fallacies that can occur from quantification.

Commentary

Meroe's and Varenne's reports call attention to the potential harmful impacts of the fallacies and influences that contribute to purpose drift, as well as to the possibility that purpose drifts are systemic, mapped onto existing structures or hierarchies of privilege. Mendoza-Denton's critique of tests with established *a priori* differences is particularly important when considering purposes such as educational management or student selection. Varenne's report calls us to consider what the role of the state should (or should not) be in anticipating and regulating purpose drift.

Cauce and Gordon's piece calls into question the extent to which particular assessment purposes supply the conditions for agentic expression. Likewise, the piece leads to consideration of whether particular kinds of "purpose drifts" limit agency or, as Mendoza-Denton discusses, trigger mechanisms of self-protection by students and teachers — as well as administrators, considering Dixon-Roman and Gergen's discussion of cheating in high-stakes environments. Like Linn, Ho provides a substantive critique of "value-added" evaluation of teachers.

Mislevy, Robert (2012). *Four Metaphors We Need to Understand Assessment.*

Findings

Needed for expert and popular discussions of assessment are a more systematic framework and more sophisticated vocabulary for organizing and distinguishing concepts underlying the purposes, designs, and uses of assessment. The four metaphors that can provide that framework and vocabulary are assessment as practice, as a feedback loop, as evidentiary argument, and as measurement. Four secondary metaphors also help provide a framework: contest, engineering, exercise of power, and inquiry.

- From a socio-cognitive perspective, conceptualizing assessment as *practice* means to consider habitus, or "the interplay between students' individual resources and targeted interpersonal LCS (linguistic-cultural and substantive) patterns" (p. 7). In what practices do we expect students to engage that evidence their capabilities?
- From a situated perspective, the metaphor of the *feedback loop* considers relationships of cause-effect or input-output, intended purposes of assessment, and the question of "who's using what for what." It also considers the effects of assessment and how it informs actions.
- The *evidentiary argument* frame considers assessment from the point of view of using data to make claims about student capabilities, where the claims are justified from established warrants or generalizations and are weighed against alternative explanations for particular outcomes.
- The *measurement* frame encompasses the quantitative tools that enable statistical inferences to be made about student capabilities, and that allow for analysis of patterns in assessment practices and outcomes. The author challenges the commonsensical conflation of "measurement" with quantification, and its oversimplification as only scores and reliability coefficients.

Each of the metaphors work in concert to help conceptualize assessment design: measurement provides the data and the warrants, the source of reasoning from which to make claims (p.26) and to assess students' marshaling of their own resources to engage in determined

practices that evidence their capabilities (p.19). Assessments also structure practices at various levels: system, classroom, and individual (p.9).

Commentary

While Mislevy mentions validity in his section on evidentiary argument, Ho's paper discusses validity as purpose, use, and interpretation. How do we synthesize these frameworks? While Mislevy seems to be crafting the argument for the consideration of these metaphors in concert, an issue brought up by Gordon Commission Fellow Juliette Lyons-Thomas is how these four metaphors may function as perspectives that can come in conflict. For example, while teachers may see tests as feedback, students may experience testing as an exercise of power (see Gergen and Dixon-Roman, p. 5). Varenne's concept of ethno-assessments suggests that we analyze the kinds of metaphors that people derive from their experiences being given and then giving assessments. The interface between those commonsense notions and those of psychometric expertise could be a fruitful space of further inquiry.

Kaestle, Carl (2012). *Testing Policy in the United States: A Historical Perspective.*

Findings

Assessment in schooling predates the implementation of IQ testing and achievement standardized testing that became widespread in the 1920s. Such assessments included oral examinations as part of everyday pedagogical practice and exhibitions, or performance assessments. The conditions for the emergence of testing included: urbanization and the practice by urban schools of using testing to evaluate teachers; Progressive era values of efficiency and scientific management; the rise of managerial class; massive immigration into the United States; and the power of the economic as metaphor for social life. The author asserts that the movement toward tracking predated the IQ, and the widespread acceptance of hereditarianism in the U.S. did not depend fully on the IQ.

Early critiques of IQ testing impacted policy by facilitating the transition from aptitude to achievement tests. Critiques shone a light on racial disparities, but ironically helped to increase the importance of testing. Epistemologically, these early critiques still held onto belief in scientific empiricism and the fairness of meritocracy.

The chief factors that support multiple-choice standardized testing include: the economy, cultural conceptions about schooling, the scientific "tradition" of testing (scoring, norms, reliability, validity), facilitation for accountability and student placement purposes, and widespread use. The chief ideologies throughout history that underlie the support of testing include: hereditarianism; technocratic segmentation (or the concept that education could be broken down into pieces); scientific management and metaphors of manufacturing; Connectionism, as developed by Thorndike — whereby "learning is a formation of a bond between a stimulus and response, bonds are formed by exercise (practice) and effect (reward)" (p. 20); and egalitarianism or equity, especially after the publication of the Coleman Report to identify underperforming schools.

The push for authentic and performance-based assessments in the accountability era have a problem gaining traction because of increased cost, increased complexity, challenge to existing frameworks, and the move toward "formative" as opposed to a focus on "summative."

Recommendations

- Implementing alternative forms of assessment requires acknowledgement of the key factors underlying the support of multiple-choice standardized testing, particularly from equity advocates.
- Align assessment with the scholarly findings about learning, and emphasize formative aspects of assessment.
- Given the history of testing, proponents of authentic and performance assessment should craft a narrative about the need and significance of these assessments that can be understandable and palatable to the general public. Additionally, they should explain how these assessments both fulfill the role of accountability and mitigate the harmful impacts of other types of assessment.

Commentary

Alongside the reports of Meroe, Dixon-Roman and Gergen, Linn, and Hill, Kaestle's piece allows for a juxtaposition of the historical development of testing with the emergences of meritocratic and democratic philosophy, the positivist paradigm (and challenges to it), modes of accountability, and technological shifts. [See the timeline in Table 1.] To this, a political-economic approach might add an account of how the historical development of education and testing has reflected changes in the U.S. economy — e.g., shifts from agriculture to agribusiness

to manufacturing, and then to service and informational economies (see Holland, et al.). This could provide insight into the intersection of assessment with histories of racial and class-based educational segregation. As the authors intend, the juxtaposition of these histories provides an understanding of the ideological and institutional supports for current testing practices, as well as opportunities for transforming practice.

Hill, Clifford (2012). *Assessment in the Service of Teaching and Learning.*

Findings

Multiple-choice tests have built-in constraints when assessing reading comprehension. Often in large-scale writing assessments, the rubrics developed for evaluating student writing are not appropriate to the skill level of the student; or, higher scores are linked to factors outside of writing, such as length. Alternative approaches to the use of multiple-choice standardized tests for selection and certification include the digital testing and digital project models developed at Teachers College.

A digital testing model based on a Pacesetter program in the United States produced more authentic assessment by using: 1) a grounded constructivist approach, whereby students were given the resources they needed, as well as tools with which they were familiar (especially in a digital age); 2) the construction of test activities instead of "items"; 3) holistic evaluation instead of a judgment based on *a priori* assumptions; and 4) an adaptation of the rubric used to evaluate the International Baccalaureate. Additionally, a process log allowed for both assessment of student time management and evaluation of the test design.

In terms of the digital project model used in China, the integration of the project with curriculum and instruction made the model successful. It also produced an authentic assessment by incorporating task guidelines and samples, step-by-step assistance, and space for peer and instructor feedback or social interaction.

Recommendations

- Use digital archives and digital assessment for the purposes of certification (such as granting high school diplomas) and selection (such as college admissions).
- Conduct periodic assessment activities throughout a school year, and have external evaluators use random selection for accountability purposes.

- Two major problems with digital assessments — authenticating students' work as their own and the costs of paying evaluators — can be addressed by: 1) using a feedback process in which student work is viewed in stages and 2) use of automated scoring and sampling techniques.
- Germany's exit vocational exams might provide a good example for the Gordon Commission to consider.

Commentary

Like Kaestle, Hill provides historical context to the development of testing in Europe and the United States. Though Hill argues that multiple-choice standardized testing in the U.S. is rooted in egalitarianism (e.g., in the writings of Thorndike), Kaestle describes the hereditarian roots of IQ testing in the U.S., highlighting figures such as Lewis Terman, who espoused views of the permanence of racial inferiority and championed educational segregation (Blanton, 2003). Nevertheless, Hill provides an example of assessment that adapts to 21st century forms of communication, and adopts a constructivist and empowerment framework, as called for by Varenne, Dixon-Roman and Gergen, and Cauce and Gordon.

Discussion

According to Kaestle, the role of history is that "it can trace the origins of important ideas and explore the reasons why some ideas get deeply embedded in educational practice and in public beliefs, and why some others have not done so" (pgs.45-46). The ten papers reviewed here signify the importance of identifying "traditions" of thought that structure educational practice (Dixon-Roman and Gergen, 2012) [see Table 2.], particularly embedded assumptions that provide the basis for "misrecognizing" educational contexts, outcomes, and stakeholders (Meroe). Additionally, historical precedents can provide insight into current educational practice, as well as suggest or reveal the possibility of alternatives, such as pre-IQ use of performance and exhibitory assessments.

One aspect of historical and social-epistemological work is to recognize how particular paradigms change over time and become rearticulated in ways that can counter or escape the negative impacts of earlier versions; e.g., nonhereditarian uses of measurement (Mislevy, 2012). Importantly, identifying tensions can contribute to understanding ideologies that contribute to the perpetuation of social injustices and birth privileges, the development of "toxic environments" in schools, and the proliferation of deficit-thinking models (Meroe, 2012; Mendoza-Denton, 2012;

Varenne, 2012). On the other hand, knowing the appeal of assessments to those seeking or promoting equity can help explain the resilience of particular paradigms (Kaestle, 2012).

Given that Dixon-Roman and Gergen cite Kuhn's (1962) work on paradigms, it might be instructive to review Kuhn's conclusions about the conditions necessary for a paradigm shift in "normal science."[1] For Kuhn, the conditions for a revolution and shift begin with a crisis that challenges the efficacy of a widely accepted paradigm (p. 145), and with the emergence of a competing paradigm. Though this latter paradigm appropriates the language of earlier paradigms, its underlying framework constitutes a significant break (pgs. 148-149). Because scientific knowledge is socially produced, a paradigm shift requires the "conversion" of a community of scientists to a new paradigm based on elements of persuasion, elegance, and "future promise," ultimately leading to changes in "professional allegiances" (pgs. 151,155,158).

Applying Kuhn's conceptualization of paradigm shift to assessment, Gergen and Dixon-Roman, Varenne, as well as Bereiter and Scardamalia, and Behrens and DiCerbo suggest that the current 21st century moment, characterized by globalization and advances in communication and informational technology, requires and demands changes in assessment paradigms and educational practices. The accumulation of evidence on the harmful impacts of traditional testing in high-stakes environments, particularly cheating scandals (Gergen and Dixon-Roman, 2012), as well as critiques of the "inadequacy" of current validity models (Ho, 2012) signal a potential crisis for educational psychology. The development of postmodern, digital, and performance forms of assessment, as well as participatory evaluative techniques represent the development of alternative constructivist paradigms that can compete with "traditional" testing paradigms (Dixon-Roman and Gergen, 2012; Hill, 2012; and Kaestle, 2012).

With assessment, as the papers from Kaestle and Ho indicate, allegiances to a paradigm must shift not only among professionals in the field of educational measurement, but also among policymakers, industry executives, and the general public. For Kaestle, this shift might occur if the right justification and cost-benefit analysis for new paradigms are clearly articulated. Varenne's attention to ethno-methodologies and the increased move toward home schooling as an alternative to test-saturated public (and private) schools indicate social pressures occurring from the "bottom-up," practices which might outpace experts and policymakers. Hakuta's

[1] It should be noted that Kuhn distinguished science from social science and medicine, particularly marking the "insulation" of scientists and their puzzle-solving from social and political problems.

imagined future impact of the Common Core State Standards initiative on English Language Learners suggests that, perhaps, we are headed for paradigmatic reform rather than "revolution," pointing to the limitations on paradigm shifts caused by conceptions of progress, the stability of patterns of practice, and, importantly, the lack of material and technological resources (pgs. 6, 11-12). The vision for a future of assessment that is dialogic, multivocal, participatory, agentic, and nondiscriminatory requires a change in a *network* of multiple paradigms or "traditions" of thought, accepted practices, and as Bourdieu (1977/1972) suggests, the historically entrenched material conditions that made those practices and paradigms possible.

The material conditions of education form another common theme throughout the 10 papers discussed in this review. Meroe directs attention to symbolic and structural resources: where the former include cultural capital, teaching quality, and social capital (such as parental advocacy), and the latter include school funding based on local tax bases, tracking, course offerings, and instructional support. Varenne points to the reproduction of birth privilege through state-regulated certification and selection processes, while Mendoza-Denton describes discriminatory school environments. In their work, Gergen and Dixon-Roman explore the tension whereby ostensibly objective assessment may be experienced by communities as "prejudice in action" (p.5). Additionally, Cauce and Gordon call for understanding the conditions necessary for the expression of agency, including access to resources and degrees of freedom and unfreedom. Allison Davis' work revealed biases in assessment toward the working class (Kaestle, 2012), and Hakuta's work highlights testing and conceptual biases against English Language Learners.

These material conditions factor into the kinds of claims that can be made about students' performance. For example, considering the work of Kozol (1991, 2005), there are students in the United States who are asked to take tests in dilapidated, rat-infested, or poorly-ventilated buildings, or with the physiological burdens of hunger, sickness, or pain. Can the types of assessments examined and imagined by the reports recognize and/or mitigate these conditions, rather than leaving them "ignored, mystified, disavowed" (Meroe, p. 5)? Perhaps, situative assessments of students, digital archives, and participatory evaluations of schools can be paired with (and account for) assessments of agency, school climate, and structural inequalities. As Cauce and Gordon point out, college scholarship assessments already consider students' experiential contexts.

Since material conditions of schooling can be impacted *by* assessment practices (Gergen and Dixon-Roman, 2012; Ho,2012; and Mislevy, 2012), one proposal for addressing unintended consequences, harmful impacts, and purpose drifts is to develop an accountability system for assessments that mirrors reviews of human subjects research. Such reviews conform to federal guidelines, yet are conducted by local boards, and yearly updates and impacts must be provided. Perhaps, such a system could lead to the kind of proactive advocacy for equity and ethics that many authors in this set of papers argue needs to come from the testing industry itself.

References

Other Commissioned Papers (2012)

Behrens J.T. & DiCerbo, K.E. (2012). *Technological implications for assessment ecosystems: opportunities for digital technology to advance assessment.*

Bereiter, C. & Scardamalia, M.(2012) *What Will It Mean To Be An Educated Person in Mid- 21st Century?*

Hakuta, K. (2012) *Assessment of content and language on the heels of the new standards: Challenges and opportunities for English Language Learners*.

Linn, R.L. (2012) *Test-Based Accountability*.

Other References

Blanton, C. K. (2003). *From intellectual deficiency to cultural deficiency: Mexican Americans, testing, and public school policy in the American Southwest, 1920–1940. Pacific Historical Review, 72* (1), 39-62.

Bourdieu, P. (1977). *Outline of a theory of practice*. Richard Nice, trans. Cambridge: Cambridge University Press.

Harding, S., ed. (1993) *The "racial" economy of science: Toward a democratic future.* Bloomington: Indiana University Press.

Holland,Nonini, Lutz, Bartlett, McGlathery, Gulbrandson, and Murillo (2007). *Local democracy under siege: activism, public interests, and private politics*, New York: New York University Press.

Kozol, J. (1991). *Savage inequalities: Children in America's schools.* New York: Harper Perennial.

Kozol, J. (2005). *The shame of the nation: The restoration of apartheid schooling in America.* New York: Three Rivers Press.

Kuhn, T. (1996 [1962]). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Moses, R.P., & Cobb, C.E. (2001) *Radical equations: Math literacy and civil rights*. Boston: Beacon Press.

Table 1. Timeline of Selected Events Related to Assessment Across the 10 Papers

551-470 BC	Confucius
300S	Chinese Imperial Examinations developed
700S	Chinese Imperial Examinations largely implemented
1600S	Conceptions of democracy developed in Western Countries
1776-1799	Revolutionary period in U.S. and France
1788	Arbitur exam introduced in Germany
1808	Baccalauréat exam introduced in France
1812	German secondary examinations
1830-1842	Comte develops theories of positivism
1871	German examinations expanded beyond Prussia
1886	Galton publishes work developing the concept of regression
1888	Classical Test Theory emerges
1892-1900	Pearson develops the concept of correlation
1904	Spearman develops the concept of the *g* factor of intelligence and factor analysis
1908-1911	Binet develops scales for determining students' "mental age"
1911	Goddard translates Binet's scales into English
1911	Principles of scientific management, Frederick Winslow Taylor
1913	Bobbitt applies the manufacturing metaphor to school administration
1916	Lewis Terman introduces Stanford-Binet test to U.S., popularizes the term "intelligence quotient"
1916	Publication of Saussure's (1906-1911) lectures, *Course in General Linguistics*
1916	College Board administers subject area exams for college admission
1917	Yerkes and the use of IQ on army recruits, Army Alpha/Beta IQ tests
1922	Walter Lipmann critiques IQ in the *New Republic*
1926	College Board introduces the SAT, developed by Carl Brigham and modeled after the army IQ tests
1929	Mannheim's *Ideology and Utopia*, an intellectual precursor to social epistemology
1936	Machine scoring of exams developed by IBM

Table 1. Timeline of Selected Events Related to Assessment Across the 10 Papers

1940s	High school attendance rose to 73%, up from 32% in 1920
1945-1953	Allison Davis critiques test bias against working-class students
1946	Stevens develops measurement scales, which spawn representational models of testing; key rupture with classical views of measurement
1950s	Post-war shift away from hereditarianism and IQ testing
1950s	Launch of Sputnik by the Soviet Union sparks educational reform in the U.S.
1962	Thomas Kuhn's, *The Structure of Scientific Revolutions* is a key moment in social epistemology
1964	Civil Rights Act is passed
1965	R.F. Kennedy's amendment to the ESEA requires testing of recipients of Title I funds
1966	Coleman Report; key watershed event for school accountability (methods and conception)
1967/1976	Publication of Derrida's *Of Grammatology*
1968	Habermas' *Knowledge and Human Interests* presents an ideological critique of the assumption that scientific knowledge is value neutral
1968	Arthur Jensen revives the IQ debate
1969-1970	National Assessment of Educational Progress initiated (moves to ETS in 1982, federal oversight in 1988)
1970s	Minimum Competency Tests (increase from 2 to 34 states from 1973 to 1983)
1983	Publication of a *Nation at Risk*, key moment in the movement for standards-based reform
1990s	Shift to Standards Based Reform, from external to state-mandated testing
1994	Publication of the *Bell Curve by* Herrnstein and Murray
2001	NCLB is passed
2009	Common Core State Standards Initiative
2010	SBAC and PARCC receive federal funds to develop assessment aligned with Common Core

Table 2. Summary of the Network of Discourse, Paradigms, and Traditions of Thought in the 10 Papers			
Social Difference/ Explanatory Schema	Foundations of Assessment	"Critical" Paradigms	Visions of Society and Social Change
Hereditarianism Market-Logics Social Darwinism Ideological and Power Analyses	Positivism Measurement Connectivism Evidentiary Argument Accountability/Influence Input-Output Feedback	Ideological and Power Analyses Linguistics and Deconstruction Social Epistemology Postmodernism Constructivism Practice, Interaction & Agency	Meritocracy Democracy Neoliberalism Redistributive Justice Collectivism Individualism

Summary of Recommendations

Conceptual Changes

- Redefine the purpose and role of education as producing and developing agency (Cauce and Gordon, 2012).
- Reframe from the question of biased or unbiased testing to one that considers threatening or nonthreatening environments (Mendoza-Denton, 2012).
- Challenge the dichotomization of the "educated" and "uneducated," that ultimately imposes deficit-models upon those with lower levels of formal education (Varenne, 2012).
- Reconsider the need for state control over educational certification.
- Replace the positivist paradigm with a constructivist one that accepts multivocality, pluralism, and particularism (Dixon-Roman and Gergen, 2012).

Institutional and Structural Changes

- Look to other countries, such as Finland, for examples of how to balance assessment, governmental support, and a collective ethos for educational systems (Meroe, 2012).
- Schools should promote an incrementalist view of intelligence, and create nondiscriminatory environments by valuing social identities and creating a sense of belonging (Mendoza-Denton, 2012).
- The testing industry also could promote an incrementalist view of intelligence, reconceptualize instruments without *a priori* group differences, advocate for the end of tracking, and relationally diversify itself as a field (Mendoza-Denton, 2012).

- Use an analysis of ethno-assessments to imagine alternative routes to the granting of privilege (Varenne, 2012).
- The state should protect the right to education and not devolve its responsibility to provide free and public education to all. However, the state should not use school-based assessments and certification (or degrees) as a means to grant career privileges. While it is reasonable to ask for assessments for selection and granting privilege, the school does not have to be the site for those assessments (Varenne, 2012).
- Take a proactive role in challenging the fallacies that can occur from quantification (Ho, 2012).
- Use digital archives and digital assessment for the purposes of certification (such as granting high school diplomas) and selection (such as college admissions) (Hill, 2012).
- Conduct periodic assessment activities throughout a school year, and have external evaluators use random selection for accountability purposes (Hill, 2012).
- Consider Germany's exit vocational exams as a model. (Hill, 2012)

Changes in Assessment

- Counter the ways that testing can "crowd out self-reflection, forethought, and creativity." (Cauce and Gordon, 2012)
- Develop holistic assessments of agency, such as portfolios (Cauce and Gordon, 2012).
- Consider the assessment of collective agency (Cauce and Gordon, 2012).
- Leverage what we know about everyday assessments to reorganize school-based assessments, considering questions such as who, what, when, and to what effect, as well as the "life-changing" stakes of assessments (p.22). (Varenne, 2012).
- Propose assessment that is dialogic, multivocal, constitutive of multiple criteria, formative, and professional-developmental for teachers and administrators (Gergen and Dixon-Roman, 2012).
- Use constructivist assessments, including empowerment and dialogic (or participatory) evaluation, appreciative evaluation, and socio-cultural, situative assessment (Gergen and Dixon-Roman, 2012).
- Keep standardized testing systems, but instead of mandating testing, expand the availability and kind of testing to communities, training them in participatory evaluation (as a form of accountability) (Gergen and Dixon-Roman, 2012).

- Align assessment with the scholarly findings about learning, and emphasize formative aspects of assessment (Kaestle, 2012).
- Develop more authentic assessments using digital testing and digital project models (Hill, 2012).

Ethics

- Emphasize to youth the links between effort, goals, and fulfillment (Cauce and Gordon, 2012).
- Take responsibility for promoting education for all as a social good (Meroe, 2012).
- Counter individualism with collective democracy and (re)distributive justice (Meroe, 2012).

Methodological Interventions

- "Critical recognition" as a methodological tool, which includes 1) historiography of concepts (social epistemology), 2) accounting not only for the "shadows" and historical consequences, but also 3) mapping the distribution of resources and recognition (Meroe, 2012).
- Conduct an inventory of the "ideological investments" supporting practices of measurement and assessment (Dixon-Roman and Gergen, 2012).
- Validity frameworks need to pay more attention to differentiating assessment of learners and learning, as well as differentiating that of teachers and teaching (Ho, 2012).
- Clearly define the targets for validation (Ho, 2012).
- "Understand the structural incentives and historical precedents" for purpose drift, as well as develop the means for anticipating and stemming purpose drifts (Ho, 2012).
- Use the four metaphors of assessment as practice, feedback loop, evidentiary argument, and measurement to develop a more systematic framework and more sophisticated vocabulary for organizing and distinguishing concepts underlying the purposes, designs, and uses of assessment (Mislevy, 2012).
- Implementing alternative forms of assessment requires acknowledgement of the key factors underlying the support of multiple-choice standardized testing, particularly from equity advocates (Kaestle, 2012).
- Given the history of testing, proponents of authentic and performance assessment should craft a narrative about the need and significance of these assessments that can be understandable and palatable to the general public. Additionally, they should explain how these assessments

both fulfill the role of accountability and counter the harmful impacts of other types of assessment (Kaestle, 2012).

- Two major problems with digital assessments — authenticating students' work as their own and the costs of paying evaluators — can be addressed by: 1) using a feedback process, in which student work is viewed in stages; and 2) use of automated scoring and sampling techniques (Hill, 2012).