The Gordon Commission
on the Future of Assessment in Education

Work in Progress

ISSUE
#2

Work in Progress – Issue #2

Work in Progress provides periodic briefings concerning the ongoing work of The Gordon Commission on the Future of Assessment in Education. The substantive basis for the work of the Gordon Commission is its Knowledge Synthesis Project, which consists of about two dozen review and synthesis papers commissioned by the Commission. Titles and authors of these papers are available in *Work in Progress Issue #1*, which may be accessed through the Gordon Commission website: http://www.gordoncommission.org/rsc/pdfs/work_in_progress_issue_one.pdf

In the conduct of the Commission's work, the several commissioned papers are the focus of study and digest by the six Gordon Commission Fellows, whose task has been to generate nominations for findings and recommendations to the members of the Gordon Commission. The Commissioners who also are reading these papers will be informed by the work of the Commission Fellows as Commissioners arrive at consensus concerning the findings and recommendations of the Commission. In this issue of *Work in Progress*, you will learn about the *Emerging Voices in the Future of Education Assessment: The Gordon Commission Fellows* — who they are, the way in which they did their work for the Commission, and what this fellowship has meant for each of them.

For this issue, we also have asked Robert Mislevy to prepare *Some Thoughts on Terminology* to clarify the meanings of several terms that are constantly used to refer to assessment and related activities colloquially and professionally. His glossary has not been vetted by the Commission, but it is useful in understanding the connotations assigned to terms commonly encountered as we try to understand assessment in education.

Finally, in our section *Abstracts for Selected Papers in Progress*, we have included summaries of the papers related to purpose, metaphors, evidence and what assessment must do. This set of papers written by Andrew Ho, Robert Mislevy, Joanna Gorin and Randy Bennett touch on issues that are pivotal to the work of the Commission.

The members of the Gordon Commission agree that the purpose of an assessment is of defining concern, and that the purposes are served by different ways of thinking about assessment (metaphors). There is consensus concerning the view that assessment is a form of evidentiary reasoning involving the production of knowledge based upon evidence. There may be less agreement, but universal concern with the perennial challenges to assessment and what this human enterprise is called upon to do. Our work on these issues continues as work in progress.

Edmund W. Gordon
Chairman
Gordon Commission on the Future of Assessment in Education

Chairman Gordon also is the John M. Musser Professor of Psychology–Emeritus, Yale University, and the Richard March Hoe Professor of Psychology and Education–Emeritus, Teachers College, Columbia University.

Emerging Voices in the Future of Education Assessment: The Gordon Commission Fellows

Curtis Malik Boykin, Teachers College, Columbia University
Ernest Morrell, Teachers College, Columbia University

Led by Dr. Edmund W. Gordon, an endowed emeritus professor at Yale and Columbia Universities and the founder of the Institute for Urban and Minority Education (IUME), the Gordon Commission on the Future of Assessment in Education brings together leading scholars to consider what educational assessment will look like — and what it should be capable of doing — now and through the middle of the 21st century. The Commission's authors are distinguished scholars in the fields of education sciences, psychometrics and public policy, and are thoughtful leaders in the arena of public affairs.

Complementing the work of the Commission's authors is the work of the Gordon Commission Fellows, a dynamic group of six emerging pre- and post-doctoral scholars in the fields of the learning sciences, anthropology, psychometrics, the sociology of education, and education technology. These Fellows were assembled to analyze and identify emergent themes, critical innovations, similarities and distinctions, and ultimately synthesize the knowledge produced across the body of the commissioned papers in brief papers of their own. It was Dr. Gordon's opinion that the work of the commission's experienced authors should be complemented by a younger generation of scholars who, in their ongoing dialogue with him and in their syntheses of the more than two dozen papers, would add new life and new ideas to the project. During their work together over the spring and summer, each Fellow selected overlapping cross-sections of the papers to critically analyze and present for a series of Fellows-led group discussions, all under the tutelage of Commission Chairman Dr. Edmund Gordon and Dr. Ernest Morrell, the current Director of the Institute of Urban Minority Education (IUME) at Teachers College, Columbia University.

The meetings of the Gordon Fellows alternated throughout the spring and summer months between IUME in New York City and the Pomona, New York-based CEJJES Institute, a cultural, educational and research foundation dedicated to improving the educational and social conditions for all disenfranchised people. In addition to Drs. Gordon and Morrell, the Fellows Meetings were co-facilitated by the Executive Officer of the Commission, Paola Heincke, and attended by project consultants E. Wyatt Gordon, Emile Session, Curtis Malik Boykin and IUME Assistant Director Veronica Holly.

These Fellows Meetings provided a platform for the six emergent scholars to put the papers into conversation with each other, with their peers and with their mentors to create a rich and dynamic analysis. The Fellows Meetings have culminated in the production of six synthesis papers, authored by the Fellows providing their analyses, recommendations, major findings and perceived implications for practice and instrumentation development policy.

Meet the Gordon Commission Fellows

Following is a brief biographical sketch of each of these dynamic young scholars, but we thought we would also ask them to share, in their own words, what participating in the Gordon Commission has meant to them personally and professionally. Therefore, following each of the short sketches is a quote, written by the Fellow, that articulates his or her views on the work of the Commission and the experience of being a Commission Fellow.

Keena Arbuthnot received a Ph.D. in Educational Psychology from the University of Illinois at Urbana-Champaign, specializing in Psychometrics/ Educational Measurement, Applied Statistics and Program Evaluation. She holds a M.Ed. degree in Educational Psychology and a B.S. degree in Mathematics. In 2005, Dr. Arbuthnot became a Lecturer on Education and a Post-doctoral Fellow at the Harvard Graduate School of Education. She is currently an assistant professor at Louisiana State University in the Department of Educational Theory,

Policy and Practice. Dr. Arbuthnot conducts research that addresses issues such as the achievement gap, differential item functioning, psychological factors related to standardized testing performance, stereotype threat, and mathematical achievement and African-American students. She also is a former high school mathematics teacher.

Dr. Arbuthnot's Reflections:

"Being a Gordon Commission Fellow has been a rewarding and worthwhile experience for me. First, it gave me the opportunity to work and network with other scholars across the nation. The discussions and dialogue regarding the future of assessment was interesting and eye-opening. The diverse backgrounds of the other Fellows presented a very fascinating exchange of ideas. During our meetings, Dr. Morrell's leadership skills were very effective. He was able to keep the group on task, while allowing everyone to express their opinions concerning various topics. Lastly, the Fellowship allowed me to work with Dr. Edmund Gordon and discuss with him the future of assessment. Given his experience in the field of assessment spanning several decades, his feedback and insight was remarkable. His comments and discussion pushed me to think outside of the box and have been instrumental in helping me to shape my future research agenda. In sum, the Gordon Commission Fellows program was a success and I am very happy that I was able to have this experience at this stage in my career. It is my hope that the Gordon Commission will be instrumental in shaping the future of assessment."

Juliette Lyons-Thomas is a third-year doctoral student in the Measurement, Evaluation, and Research Methodology (MERM) program at the University of British Columbia. Her current research focuses on think-aloud protocols as a validation method in educational assessment. Her interests also include accountability in education, validity and cross-cultural assessment. Juliette received her M.A. from New York University in Educational Psychology, specializing in Psychological Measurement and Evaluation, and her B.Sc. from McGill University in Psychology.

Ms. Lyons-Thomas' Reflections:

"As part of this project, I've had the privilege of getting a first look at papers that were written by leading education researchers, many of whom I have admired for some time. Beyond that, the project has given me a platform to connect with some of these esteemed thought leaders; this opportunity has been so much appreciated because, as a graduate student, approaching these individuals can otherwise be very daunting. Another benefit of participating in this project was being able to discuss the concepts with people from other fields. I felt that the varying perspectives were always valued in our meetings, and I was able to see problems from a point of view that I would not have otherwise considered. After spending time with the other Fellows who have truly become role models, I always came back from our meetings feeling inspired to improve my contribution to the Commission, as well as further my own work that is separate from the Fellowship. On more than one occasion, I have been able to connect with other Fellows in between meetings to discuss my doctoral research and I have received invaluable feedback as a result. Professor Morrell facilitated stimulating discussions during our meetings and often guided the group on particular points that could be unpacked, teased apart or elaborated upon. He consistently provided experienced leadership to our group of early-career academics. Finally, having the privilege to speak directly with Professor Gordon, and hear his thoughts and reactions to our discussions, is a memory that I'll carry with me for the rest of my academic career. My experiences with the project have been exclusively positive, and I'd very much like to stay involved with IUME and the Gordon Commission after this project comes to a close."

Jordan Morris is a second-year doctoral student in the Social Welfare program at the University of California, Los Angeles. She received her B.A. in Psychology from the University of Maryland, College Park, and her Ed.M. in School Psychology and Education Policy from Teachers College, Columbia University. Her research interests include child and adolescent development, critical media literacy, and race and schooling.

Ms. Morris' Reflections:

"Being part of the Gordon Commission has been a very valuable experience for a number of reasons. First, Professor Morrell provided excellent leadership in guiding us through the process of synthesizing and pushing us to appreciate the many Commission

papers. Second, this experience readdressed the many issues that frustrated me with educational assessment in the past, but also showed promise that these issues are going to be addressed in the future with the wide-open possibilities that new technologies bring to education. Third, collaborating with a diverse team seemed like being part of the solution education wants, needs and desires. Finally, working with Dr. Gordon was inspiring with his breadth of knowledge, and productively challenged my view on assessment. Overall, this experience has helped me academically and professionally to rethink some of my own work moving forward."

Catherine Voulgarides is a fourth-year Ph.D. student in the Sociology of Education program at New York University, where she currently works as a research assistant at the Metropolitan Center for Urban Education under the leadership of Dr. Pedro Noguera. At the Center, she has worked on and assisted with the Technical Assistance Center on Disproportionality in Special Education. Before joining the Center, she worked for the AmeriCorps Vista project in Phoenix, Arizona, coordinating and developing ESL programs for recent immigrant parents in the Phoenix school system. She holds a B.A. in Economics and is a graduate of McGill University in Montreal, Canada. She also holds a MST in Special Education from Pace University in New York City, and taught middle school special education for several years in Washington Heights. Her research interests are centered on the intersection between federal disability legislation and racial and ethnic disproportionality.

Ms. Voulgarides' Reflections:

"Through my experiences as a Fellow in the Gordon Commission, I have been able to more clearly understand how the current testing regimen, centered on accountability and improving student outcomes, stifles some of the advancements being made in assessment and measurement. I believe that the work of the Gordon Commission creates a much-needed space for influential scholars to explore and imagine a more equitable and informative assessment system that places learning for the sake of learning at the center of education, rather than focusing on learning to meet a benchmark or policy mandate. I think that being paired with five other Fellows from varying academic disciplines allowed me to explore the future of assessment in a nuanced and dynamic way. I was able to not only see how my background and training as a sociologist of education relates to assessment, but I also grew as a scholar as I engaged with the perspectives of the other Fellows. I feel lucky to have been part of this experience, and think that the papers written for the Commission and the recommendations that will follow must be seriously grappled with by not only the assessment community but also by other academic disciplines, policymakers and the public at large."

Amanda Walker Johnson received both a Ph.D. and an M.A. in Anthropology (Sociocultural) from the University of Texas at Austin's African Diaspora Program. In 2004, she served as both a research associate for the Research and Evaluation Division at the Intercultural Development Research Association in San Antonio, Texas, and as an assistant instructor in the Department of Anthropology at the University of Texas at Austin. In 2005, Dr. Johnson was hired as an adjunct faculty member in the College of Humanities, Arts, and Social Sciences at the University of the Incarnate Word in San Antonio, Texas. In 2006, she was hired as an assistant professor in the Department of Anthropology at the University of Massachusetts Amherst. Dr. Johnson's areas of expertise include African-American anthropology; critical race theory and political economy of race in the United States; critical educational theory; feminist theories of race, body and nation; anthropology of science; and cultural and identity politics in the African Diaspora.

Dr. Johnson's Reflections:

"In our meetings, I was struck by Dr. Morrell's ability to clarify difficult points, re-encapsulate discussions and main points, synthesize several works and comments, raise critical issues and challenge us to expand our thinking. From advice on professional development to work-life balance, his leadership among the Fellows was invaluable. Coming together with the Fellows in the synthesis project for me meant more than digesting scholarship about assessment. I found in the content of the papers and the experience of the Fellowship lessons for my own roles as scholar, teacher and parent. The warmth of Professor Gordon, Paola Heincke, Dr. Morrell, Veronica Holly and IUME created an environment of being invited into a family

and provided a source of mentorship infused by authentic caring."

Sherice N. Clarke is pursuing a Ph.D. in Education at the University of Edinburgh, anticipating the award of her doctorate in the spring of 2012. Her thesis is titled *The Inclusive Museum: Understanding Adult ESOL in Museums*. Clarke currently holds a M.Ed. with a concentration in Teaching English to Speakers of Other Languages (TESOL) from the University of Edinburgh, as well as a bachelor's degree in Art History from Hunter College. Her research interests include engagement, agency, identity, classroom discourse and narrative. She currently holds a postdoctoral appointment at the University of Pittsburgh's Learning Research & Development Center. Additionally, she has been an instructor in the University of Pittsburgh's Linguistics Department, a teacher trainer at Edinburgh's Institute of Applied Language, and an English teacher and EFL Department advisor at the Sathit Bangua School in Samut Prakam, Thailand.

Ms. Clarke's Reflections:

"It has been a unique opportunity working with Professor Gordon, Professor Morrell and the Fellows synthesizing the commissioned papers and examining some of the core issues in educational assessment. In particular, the interdisciplinary nature of the group afforded the opportunity to engage with the core issues from a range of disciplinary perspectives, which has not only broadened my perspective on key issues in measurement, but also key issues in the learning sciences."

Future Plans for the Gordon Commission Fellows

Now that the synthesis papers have been completed, our goals are to share the work of the Fellows with multiple audiences via social media, conferences, colloquia, published papers and meetings with key stakeholders. This fall, the Fellows will travel to Washington, D.C., to meet with the National Academy of Education's postdoctoral and dissertation Fellows to discuss their work and its implications for the field. During this visit, we also hope to have Fellows connect with elected leaders and others who work with education at the federal level. In addition to the District of Columbia visit, we also are planning a colloquium on assessment to be held at Teachers College that features, among others, Professor Gordon and the Gordon Commission Fellows. Finally, we are submitting abstracts for upcoming conferences and outlining for possible publication manuscripts that deal with the future of educational assessment as well as the process of intergenerational collaboration that is at the heart of the work with the Fellows.

Some Thoughts on Terminology

Robert Mislevy

People use some of the central terms in educational assessment in different ways, leading at times to confusion and cross-talk. To reduce this problem within our own writings, I would like to suggest the following as starting points for a discussion of key meanings we might be able to agree on and use consistently across our chapters. If a writer has good reasons to use a different sense within his or her chapter, this can certainly be done, but the reader would be aided by a mention that the word is being used differently and the point that this use helps make. Terms within the two groups are sometimes used as synonyms, but I want to suggest distinctions that will be useful.

Assessment. An assessment is a means of obtaining information about a student's learning, capabilities, attunements, accomplishments, or intellective or other personal characteristics. The information gathered from that assessment may be used for any of a number of purposes, including such low-stakes purposes as guidance and instructional planning and higher-stakes ones related to promotion, graduation, teacher bonuses or school effectiveness. Tests are one category of assessment; so are minute-by-minute reactions in a tutoring session or a game, a long-term project like AP® Studio Art portfolio assessments, and individual evaluations by a school

psychologist. Measurement models may be, but need not be, employed.

Test. A test is a particular kind of assessment — a well-defined event, usually limited in time, usually with pre-defined goals set out for the student, usually with defined procedures for evaluation. Some sorts of scores, or qualitative characterizations (e.g., diagnoses), are usually generated, but formal measurement models may or may not be employed in generating those scores or qualitative characterizations.

Measurement. Educational measurement concerns the application of mathematical/statistical methods to the information obtained in tests or assessments. In contrast to "educational assessment," "educational measurement" has traditionally been interpreted to mean that there is some characteristic that exists in students to different degrees, which generalizes to situations outside of the test and which is evidenced by scores (or qualitative characterizations). However, the use of measurement models also can be viewed as engineering approximations, to help manage information and uncertainty, to make sense of large volumes of test or assessment data, to evaluate and improve test functioning, or communicate information obtained from tests or assessments.

There is a particular assessment configuration currently used in programs such as No Child Left Behind (NCLB) and Race to the Top that is variously called standardized testing, high-stakes testing, accountability testing and multiple-choice testing. Each of these terms, however, addresses certain features of the configuration which don't need to be linked with others. It is therefore useful to be able to make the distinctions, in order to draw contrasts among different possible kinds of assessments for various purposes.

Standardized. Assessments have myriad characteristics, and it is a design decision as to which of them should be the same for all examinees (or otherwise controlled in some way for all examinees). Timing, time limits, sources of support, task definition, type of material presented, often the test items themselves, conditions of performance, forms of responding, requirements for work products and evaluation procedures are all examples. All of these choices can, in principle, be made distinctly from one another, in order to tune an assessment to a purpose and a context.

Multiple-choice. Multiple-choice tests are distinguished by the form of responding to tasks — standardized, and in particular response that is limited to the indication of a particular "most correct" answer. Multiple-choice tests can be used in high-stakes or low-stakes tests (see below). Especially with technology-supported tests, tasks can be open-ended from the perspective of the student (such as creating a process model, filling in a representational form, building an explanation from a series of drop-down menus, or indicating troubleshooting actions in an interactive simulation), but be essentially multiple-choice — maybe with very many alternatives — from the perspective of the examiner.

High stakes. High-stakes tests are ones for which results have important consequences for someone, whether it be the student, the teacher, the school, etc. College admissions tests have high stakes for examinees, NCLB tests have high stakes for schools (but not individual students), and doctoral dissertations are high-stakes assessments that can take years and in many ways are not standardized. An assessment may be high-stakes for one group but not for others, as in the NCLB example above.

Accountability. Accountability is related to high stakes, but focused on high stakes for particular individuals or institutions that have responsibility in an educational system. Assessments used for accountability purposes nowadays often are standardized, but this isn't always so and doesn't need to be.

Purposes, Metaphors, Evidence and What Educational Assessment Must Do

Edmund W. Gordon

Purposes

When Binet was commissioned to create a test of intelligence, the purpose behind the request was to develop a procedure by which French society could identify those persons who could benefit from the then-scarce resource we call education. The history of human societies can be marked by changes in who societies have considered to be educable. As the human species has advanced in its evolution and the ideas of democracy and social justice have emerged, more and more categories of persons have been deemed to be educable. Binet's test of intelligence enabled societies to use aptitude for learning, previously developed academic ability (intelligence) and later academic achievement as the universal indicator of educability. Assessment in education became preoccupied with the measurement of one's status with reference to those developed abilities. But assessment in education does and can serve other purposes.

The first paper in this collection, *Variety and Drift in the Functions and Purposes of Assessment in K–12 Education*, by Andrew Ho, reminds us of the serious problems related to drift in the use of assessment instruments and data from the purposes for which they are designed. He addresses an issue that has been central to the work of the Gordon Commission: the multiple purposes for which assessment in education is and can be used.

A wide variety of purposes have been identified, including:

- Measurement of the status of one's developed abilities
- Inventory of one's characteristics and abilities
- Sorting and rank ordering of subjects
- Accountability
- Prediction
- Selection
- Evaluation
- Diagnosis
- Informing and improving teaching and learning

It is the last purpose in this list — "to inform and improve" — to which many members of the Gordon Commission have called attention. The Chairman of the Commission views "to inform and improve teaching and learning" as the purpose that includes and is superior to all other purposes. In his view, the other purposes listed are and should be instrumental to the production of knowledge to better inform and to improve the processes and outcomes of teaching and learning.

The assertion of this position is not to demean other purposes such as sorting, selection, placement and even accountability. But it is to argue that the investment of so much teacher and learner time in the preparation for and the engagement in assessment exercises is, in my view, justified only by the contribution that such engagement can make to better informing and improving the processes and outcome of teaching and learning themselves. If students are to spend time in preparation for assessments, the assessments for which they are preparing should demand the competencies that should be the goals of education. Increasingly, those goals are not co-terminus with rote memory, recognition of the right answer or regurgitation of factoids.

Now the design of assessment instruments and procedures that both inform and improve teaching and learning may require that the products meet certain criteria. We have not yet turned attention to such specification, but that work should be on the agenda of any continuation of a focus on the future of assessment. The learning sciences and emerging research in curriculum development are beginning to provide some leads. That kind of R&D will not happen by accident. Those of us who use assessment data and technologies and especially those of us who demand and pay for assessment data, as well as those of us who teach and learn, need such data in order to do our jobs well. We will need to demand that our assessment instruments and programs produce information that informs and improves teaching and learning processes and, eventually, the

outcomes of one's having engaged in these processes.

To achieve this end, we will need to bring greater balance to our national and state education policies that currently place disproportionate emphasis on accountability as the primary purpose of assessment. Our nation is making a sizeable investment of money and human capital in the pursuit of improved instruments and programs of educational assessment. Any investment in education is welcome, but that such capital should be invested in assessment instruments and procedures that better inform teachers and learners with respect to how we teach better and how we learn better. Accounting for what we have done may contribute less to that end than the production of knowledge that tells us how to do it. Ho correctly cautions against the tendency to drift from the purpose for which an instrument or procedure was designed. There are serious transgressions in contemporary policy and practice. Student achievement tests were not designed to be solitary indicators of teacher quality. Nor are they the best indicators of the ways in which a specific student learns or of the developable potential of specific students. Our instruments and procedures should be used only for the purposes for which they were designed, until they have been subjected to the R&D investigations to determine suitability for other purposes.

As we move forward, assessment in education will need to be directed at 1) the recognition of the multiple purposes for which assessment data, instruments, and procedures can be used, 2) at ensuring the empirically supported use of these for purposes other than the purposes for which they were designed, and 3) at the achievement of greater balance between our attention to these multiple purposes, so as to avoid the disproportionate and possibly distorting influence that the emphasis on accountability may have on education and on the character of the assessment enterprise itself. The enterprise may place so much emphasis on holding us accountable for meeting the common core standards that we will lose the opportunity to improve our capacity to produce well-educated people as well as to produce the actual instruments and procedures that can help us teach and learn what such standards are or should be intended to achieve. These are examples of the several issues with which the Gordon Commission on the Future of Assessment in Education is concerned.

Metaphors

In a continued effort at increasing understanding of the domain of knowledge with which we are concerned when we refer to assessment in education, Bob Mislevy was commissioned to write the paper *Four Metaphors You Need to Understand Assessment*. This is the second paper in this collection and it complements Ho's paper on purpose with its contribution of ways of thinking about assessment and just what it is that we are trying to do when we engage in assessment activities. These ways of thinking about assessment reflect the complexity of the enterprise and forecast possible directions in which the field may move as we contemplate the future of assessment in education.

Evidence

The third paper in this collection, *Assessment as Evidential Reasoning*, addresses the core processes implicit in assessment activity. What is assessment about? At its core, assessment is concerned with reasoning from evidence. Even though purposes may vary and the ideas behind the process may differ, the activity involves making logical inferences from and supporting those inferences with evidence. Joanna Gorin was asked to write this paper for the Commission as grounding for all else that we do. Assessment assumes warrants by which we make judgments that are based on evidence of phenomena that are logically related to the purpose for which the assessment data are collected.

What Assessment Must Do

At the core of the charge to the Gordon Commission is the expectation that from our work will come some direction for the field of assessment in education, given the Commission's analysis of where we are and have been, and our projections concerning where we can and should move. Randy Bennett concludes that the challenge to the assessment community is to respect its foundational principles even while applying them in new ways to meet the demands of dynamic and rapidly changing times. Some members of the Gordon Commission argue that respect for the best of what we have includes continuing

critique, disconfirmation and re-conceptualization of those principles. Guided by the work of the Gordon Commission and a distinguished career of relevant R&D, with the near future and practicality in mind, Bennett has generated 13 wise and feasible actions to guide what the field can and should do in preparation for the future of assessment in education.

Variety and Drift in the Functions and Purposes of Assessment in Education

Andrew Ho

Introduction

Validity is a quality of the interpretation and use of an assessment, rather than the assessment itself. It is based on an interpretive argument grounded in a clear statement of purpose. However, outside of the literature it is rare to see an assessment framework built with "purpose" as a central concept. Instead, assessments are used for multiple, underspecified purposes over time.

This paper reviews and adapts different frameworks for classifying the purposes of assessments — large-scale standardized K–12 general education assessments in particular. The author reflects on the forces that shape the uses an assessment is put to and the expansion of those purposes over time.

The author begins by exploring Haertel's distinction between assessment for measurement and assessment for influence. He maps the National Research Council (NRC) report *Knowing What Students Know* onto this framework and overviews Kane's (2006) chapter on validity with a focus on assessment purpose. The author highlights the tension between a focus on presentation (limiting the number of discussed purposes at the cost of accuracy) and a focus on accuracy (increasing the specificity of discussed purposes at the cost of conceptual utility).

Haertel

Haertel identifies seven purposes of assessment. A 2005 chapter by Haertel and Herman describes the rise of five purposes for assessment throughout different eras since the early 20th century. Two broad purposes are consistent throughout time: assessment for individual placement and selection and assessment to improve the quality of instruction. The Elementary and Secondary Education Act of 1965's emphasis on comparing the relative effectiveness of curricula gave rise to the use of assessments in educational program evaluation.

The National Commission on Excellence in Education's *A Nation at Risk* report marked the rise of two other purposes: using assessments to shape public perception and to focus the attention of the educational system on reform. Assessments are used both to identify underperforming students and schools and as a barometer of their success in and commitment to addressing the problem. International comparisons rise from this assessment purpose.

In a 2012 address, Haertel added two additional purposes. The first is education management via the measurement of teacher and school effectiveness in a way that supports making inferences and decisions about teachers and schools. The second is directing student effort, where assessments inform on which areas students should focus their efforts. Haertel leaves the unintended consequences of those intended purposes for other authors to untangle.

Table 1. Seven Purposes of Educational Assessment (from Haertel, 2012)

Measurement	Influence
Instructional Guidance	Directing Student Effort
Student Placement and Selection	Focusing the System
Informing Comparisons Among Educational Approaches	Shaping Public Perceptions
Educational Management	

Haertel divides the purposes into Measurement and Influence purposes. Measurement purposes rely on the information provided by test scores, while Influence purposes flow from testing independent of specific test results. Classroom standards-based assessment can direct student effort and increase attention to the desired learning outcomes regardless of a student's or class's specific scores. Haertel argues that it is insufficient to support an interpretive argument for a Measurement purpose when an Influence purpose is also essential. Attention to the validation of Influence purposes of an assessment regimen tends to reveal unintended consequences such as test score inflation and the narrowing of educational goals to the tested content.

Knowing What Students Know

The 2001 NRC report *Knowing What Students Know* identifies three purposes of assessment: assisting learning, assessment of individual achievement, and program evaluation.

Assisting learning can be understood as related to the concept of the use of formative assessment to inform instruction. In Haertel's categories, this is understood as two separate purposes: on the Measurement side, "instructional guidance" covers the use of specific test results to assist teachers in improving instruction, and on the Influence side, "directing student effort" covers the indirect impact ongoing formative assessment has on facilitating student engagement.

Individual achievement includes various kinds of summative assessments such as end-of-course grades, admission and selection assessments of postsecondary institutions, and individual scores on state accountability assessments. Many aspects of this purpose align with Haertel's more specific "student placement and selection category." The specificity of Haertel's category allows us to avoid conflating summative assessments that inform instruction with those used for individual or school accountability, or conflating formative assessments summarized for a summative accountability purpose with those used to make judgments about learning trajectories.

Program evaluation assessments include those that support aggregate scores, from small-scale research to large-scale assessments such as PISA or NAEP. This purpose aligns to Haertel's categories of "informing comparisons among educational approaches" and "educational management." Making inferences about teachers and principals instead of programs involves making a distinction between school personnel and their actions, which can be difficult. The NRC report explicitly aligns assisting learning, individual achievement and program evaluation to Measurement goals around learning, learners and programs, respectively, aligning to Haertel's "focusing the system" category. The signaling of those goals requires no test results per se, but is assumed to focus the implementation of the testing regimen.

Kane and Validity

The comprehensive treatment of validation in Kane's 2008 *Educational Measurement* chapter provides a useful practical framework for validation. Kane illustrates *trait identification*, in which Kane defines a trait as "a disposition to behave or perform in some way in response to some kinds of stimuli or tasks, under some range of circumstances." Trait labels and descriptions imply values and assumptions and make predictions and justifications that require interpretative arguments. Haertel's categories do not incorporate trait identification except to the degree that the interpretive argument for trait identification extends to student placement and selection and instructional guidance.

Kane extends trait identification to *theory development*, where relationships between traits and other phenomena are established. The incorporation of traits into regression models aligns with Haertel's "informing comparisons" purpose. Kane goes further to provide a framework for the validation of large-scale accountability programs, the purposes of which are explicitly laid out in Haertel's framework (Table 1).

The Union of Frameworks

Haertel's framework is intended primarily for standardized, large-scale achievement testing, and has incomplete applicability to classroom and formative assessment and trait and theory

development. Understanding formative assessment in this framework requires a conception of formative assessment as a process rather than a product, one that incorporates teacher and student training in assessment and feedback. From this perspective, although the measurement of learning is crucial to fulfilling the Measurement goal of "instructional guidance," high-quality formative assessment practices change classroom practice regardless of the results of the assessments themselves, fulfilling the Influence purpose of "directing student effort."

Incorporating trait estimation and theory development requires representing Measurement endeavors of "student placement" and "informing comparisons." The level of the theorized impact of the traits and models distinguishes these. Both trait estimates and student placement require inferences about individuals, while theory building and informed comparisons require inferences about the relationships at the aggregate level. Large-scale national assessments like NAEP are examples of measurement at a still higher level of aggregation. Their Measurement purpose of informing comparisons is clear, but their Influence purpose is unclear. While they are influential in shaping public perception and focusing the system, the methods by which they do so are deeply dependent on results, making those impacts more of a Measurement purpose. Fitting them to Haertel's dimensions may require the addition of an additional Measurement purpose for large-scale demographic and national comparisons.

Additional Framework Dimensions

Haertel's framework incorporates additional dimensions not described in this paper, included an assessment's primary users, the constructs being measured, linkage to the curriculum, etc. These kinds of dimensions deepen our understanding of interpretative arguments and highlight variety in assessment purposes. The author highlights the dimension of the provision of useful feedback to inform teaching and learning. Many, though not all, have this as a direct or indirect goal. In those cases, a useful dimension for classifying assessment purposes is the length of the feedback loop from the assessment event. In other words, once an assessment is given to a student, how long does it take for useful feedback from that assessment to return to that student? Feedback loops are dependent on the timeliness of scoring, reporting and decision making.

The length and impracticability of feedback loops raises the question of whether assessments not explicitly designed to provide the short turnaround characterized by formative assessment could serve a formative Measurement purpose. The more realistic theory of action for large-scale and summative assessments lies in Influence purposes, where the assessments direct efforts in response to incentive structures. This may change as technology advances live scoring and computerized testing systems.

Another dimension involves the stakes of a particular assessment-based decision or inference. Stakes are lower for assessments built for formative, academic research and national-level comparison purposes. Stakes are higher for assessments built for trait identification, selection and educational management purposes. Stakes are an important consideration because they demand rigorous interpretative arguments. "Value-added" systems require a questionable interpretation of observed and predicted scores to make inferences about the value of a teacher, but even if the interpretative argument was strong, theories of how teachers should improve (a Measurement purpose) or how incentivizing teachers via value-added scores improves student learning (an Influence purpose) have been left unspecified.

Anticipation of and Response to Purpose Drift

The second goal of this paper is to describe the tendency of modern assessments towards *purpose drift or purpose creep* — the strategic, opportunistic and relative adoption of new purposes for existing assessments. Much of the struggle with the purposes of assessment springs from the difficulty of explaining to non-academics that validity is not a property of an assessment but its use and interpretation. Validity as it is defined and defended during test development has little bearing on the responsibility of the test user to appropriately utilize an assessment. The notion that an assessment, once validated, can be used for anything is consistent with the common idea

that numbers "travel" — an idea combining a host of appealing fallacies of reasoning that allows test users to ascribe various meanings to numbers as their shifting purposes dictate. Selection tests like the SAT® are not designed to be used as components of state accountability testing, and existing tests are not designed to be fit into a statistical model for making "value-added" judgments about teachers.

If known forces cause the purposes of an assessment program to deviate from the purposes originally validated, then conventional validation approaches proposed in the assessment literature are inadequate. Validation needs to be framed proactively in anticipation of purposes to come. If publishers and policymakers don't change their practices, validation will be reduced to toothlessly scolding end users long after high-stakes, indefensible decisions have been made about students, teachers and schools. A deeper understanding of purpose drift calls for raising the standard of validation to proactively stem the anticipated drift of assessment score purpose. While purpose drift may be impossible to prevent, we know that it occurs and changes the consequences of the implementation of an assessment.

Four Metaphors We Need to Understand Assessment

Robert Mislevy

Introduction

Discussion about how to improve assessment is hampered by the lack of a common language to describe and interpret issues. People rely on their personal experience and expertise, which falls short for unfamiliar kinds and contexts of assessment. This paper's goal is to provide a guide for four metaphors or perspectives for understanding assessment:

- Assessment as Practice
- Assessment as Feedback Loop
- Assessment as Evidentiary Argument
- Assessment as Measurement

Assessment as Practice

Assessments are recurring, organized sociocultural activities in which people engage. The capabilities an assessment requires of students share some overlap with the real-world practices for which the assessment is meant to prepare them, but they are not one and the same. What students learn from an assessment often is limited to the assessment context — e.g., the ability to "do" algebra on the test does not necessarily line up with the ability to use it in any other context. Assessments occur in a wider social context, and their use influences learning and shapes the way people think about students and capabilities. This contextualization shapes the effective meaning of all the variables and inferences in an assessment system, regardless of whether this is taken into account in its design.

Assessment takes place at the level of individual actions and experiences. That activity is mediated by large-scale, long-term processes that define cultural conceptions of social institutions like marriage, narrative structures like scientific models, and semiotic systems like language. Communities and disciplines are marked by identifiable recurring clusters of themes and activities called *practices*. At the same time, individual actions and experiences are produced by small-scale, short-term processes. Neural patterns must both relate to cultural patterns and adapt to unique situations to produce successful individual-level activity.

Instruction and assessment activities share some key cultural patterns and activities with real-world situations, despite their distinct contexts. Their goal is to develop and observe the neural capabilities necessary for students to produce successful individual activity in a real-world context. We can think about assessment in terms of the interplay among the targeted cultural practice and students' internal resources that underlie their performance.

The capabilities an assessment requires share some of the capabilities of the real-world practice for which it is meant to prepare students, but the assessment is not the same as those practices. Again, what students learn often is bound to the conditions and practices in which students learn. An assessment can only signal to students what is important to learn

for the practice of that assessment. Whether those are the same as the standards of the real world is a separate question. We cannot assume that students will develop the kinds of thinking and acting we care about if our assessment and instruction doesn't authentically engage those capabilities. This is the argument for performance assessments as well as epistemic and other kinds of contextualized work.

It is more difficult to assess contextualized capabilities like "21st-century skills" than it is to assess knowledge and tasks. They are not well-defined skills that everyone will develop and demonstrate in the same ways in the same situations. They are ways of thinking and acting that manifest in different ways in different domains. Indicators of evidence must be interpreted in each individual's context and history. There is no decontextualized, standardized test of a concept like "intellective competence."

Assessments are a cultural practice. They influence how people think about learning and what they consider important to know. The "assessment as practice" metaphor makes us aware that assessments don't simply measure existing qualities in students; to a large degree, they cause our conceptions of qualities to exist. As such, they have a key role to play in reform.

Assessment as Feedback Loop

Education is about improving the capabilities of students. Everyone in an assessment system needs to understand information about student learning, from state school officers to teachers to tutors to the students themselves. The timescale and context of that information vary from perspective to perspective. Careful consideration of who needs what information, when, and for what reasons has implications for the properties and purposes of the assessment themselves.

A key insight of this perspective is that the value of assessment data is not inherent to the data, but depends on who is using it for what decisions. The same assessment can be invaluable in one context and useless in another. Therefore, high-quality assessment design looks not just at tasks, but at how to provide the best information for whoever needs it, in light of what they need to do.

One-on-one tutoring sessions are invaluable to extending student's resources in the short term and providing precise feedback to an instructor, but a video clip of the same exchange may be unintelligible to an outside observer. Classroom quizzes are standardized to the level an individual teacher needs to gather information about students, but not standardized to the level demanded by state agencies. Whether used by teachers or students, standardization is a tool that can be applied selectively to different aspects of an assessment for different purposes, not an all-or-nothing characteristic of a test. Colleges value a standardized test like the SAT because of its invariance, even though the evidence it provides about individual students is decontextualized and less in-depth than that provided by GPA.

The AP Studio Art portfolio assessment blends standardized and non-standardized features to support feedback loops at different levels of the educational system. The work that is judged is produced in thousands of individual teacher-student interactions, and through training and feedback both students and teachers are coached in defining problems and evaluating work. At the same time, the program's requirement to produce certain numbers and kinds of pieces and the centralized process for setting judgment standards make standardized ratings possible. The design balances support for student learning with summary evaluations that inform college admissions personnel.

The effectiveness of an assessment in informing feedback is reflected in the trust people place in the assessment. Jim Gee points out that video games provide active feedback and use that information to adjust challenges and respond to the player, and are viewed as effective assessments of mastery of the game. Completing an algebra course, on the other hand, requires outside assessment because that datum indicates very little about the skills the student has developed. We administer external tests for algebra because completing most algebra courses provides less information on student learning than completing a video game. At the same time, video games tell us little that is needed for college admissions. Good policy has to recognize the feedback loops that are created and served by different assessment designs.

Assessment as Evidentiary Argument

The "assessment as argument" metaphor connects the sociocognitive metaphors used so far with the systems perspective of measurement. Evidentiary reasoning helps us understand assessment in context. It provides a framework for reason, from what students do in a limited set of situations to what they can do in a broader context or should work on next. It connects the situated and practical perspectives of the practice and feedback metaphors with the technical tools of the measurement metaphor, and is proving helpful in developing educational simulations and games.

In this approach, the assessment designer asks what knowledge and skills should be assessed because of their value to society, then thinks of what performances provide evidence of that knowledge, and finally designs tasks that are able to elicit those performances. The assessment makes a *claim* about a subject supported by *data* and a *warrant* that justifies making an inference or claim based on that data. Theory and experience provide *backing* for the warrant. Explanations are qualified based on *alternate* explanations for the data, which may have rebuttal evidence that supports or refutes the alternate explanation.

Data about an examinee's performance, aspects of the task and contextual information about the examinee ground the claim an assessment makes about the examinee. Warrants tell us what is important in the assessment context, what to look for in examinee performance, and the terms of what can be inferred about examinees beyond their immediate performance. Alternative explanations involve examining what, other than the favored claim, could explain the examinee's performance.

The argument metaphor emphasizes that a claim or argument about the subject being assessed is made from the knowledge standpoint of the assessor. Context gives additional information the assessor can use to improve his or her interpretation of performance, making more precise inferences and undercutting alternative explanations. Standardization also can mitigate alternative explanations, at the cost of reducing the contextual information needed to make precise inferences. Standardization, from this viewpoint, is most useful when there is insufficient contextual evidence in the first place: a standardized test that is useful when comparing students who've completed different, unknown algebra courses is not as useful when comparing students who've taken the same specific Carnegie Learning® algebra course. The more abstract the quality an assessor wants to make an inference about, the more the assessor must take context into account. One compromise may be standardizing some aspects of a performance assessment but not others, as in the AP Studio Art exam. The evidence-centered design (ECD) framework is one example of the tools that have begun to appear to help designers craft assessments using the principles of the "assessment as argument" metaphor.

Assessment as Measurement

Measurement is the most familiar metaphor because it has been a primary theme of assessment for a century. Measurement models presume that it is possible to measure well-defined quantitative properties of students. Developments in psychology challenge this notion: instead of literally measuring traits, assessment results are now understood as patterns of evidence about capabilities in context. That said, measurement models can guide the use of assessments and data even if they cannot literally represent internal qualities of students.

Despite common wisdom, educational measurement is not synonymous with educational assessment. The variables in measurement models, such as intelligence and reading comprehension, are still considered by some to be quantifiable properties similar to force and mass. An alternate view of measurement models is metaphorical — their use is an example of model-based reasoning. The patterns of evidence we refer to as intelligence and reading comprehension arise not because of values inherent to the people being measured, but because of patterns in how people learn and their linguistic, cultural and social practices. The measurement model is not a representation of a physical truth, but a frame for reasoning about performance patterns that differ under different circumstances.

There is more to measurement models than numerical scores. The AP Studio Art program doesn't assign an "art proficiency" score; it monitors patterns of ratings across performances to identify atypical instances that signal errors or unique performances. Measurement provides a quantitative framework to augment qualitative arguments. Measurement also enables the kind of ongoing embedded assessment based on large amounts of continuous data points that is enabled by digital technology and advances in computing. Using this ability, researchers are now advancing automated scoring of complex performances, data mining from game and simulation environments, and unobtrusive psychometric modeling.

Four Additional Metaphors

We can contrast *tests as contests* with *tests as measurement*. When examinees compete against one another for scarce resources, such as jobs or college admission, incentives exist for examinees to maximize their chances. This includes preparation in ways that favor scores over learning and cheating. Test designers must take the competitive context into account, and strive to provide tests for which better learning is likely to produce higher scores. The tests need to produce reliable outcomes through transparent scoring in order to minimize cheating.

Traditional tests rely on limited, independent tasks written to fit into test specifications and specific curricula. That process is insufficient for designing assessments for interactive performances, multiple aspects of knowledge, collaborative exams, or other new ideas in assessment design. *Assessment engineering* is a new line of work that involves developing concepts and tools to improve the quality of task design for these new challenges. Embretson's cognitive assessment design system, Luecht's integrated test design/development/delivery, and Mislevy's evidence-centered design framework are examples. The *assessment engineering* metaphor is necessary for the efficient development of evidentiary argument machinery and simulation-based assessment.

Examinations are an *exercise of power* by which modern societies make visible and categorize individuals along "normal" categories. Language testing can be used as a weapon between groups in society; economists view professional licensure through the lens of barriers to entry. This metaphor is useful for describing the shaping effects assessments have on people, including in some cases the abuse of authority. Exercising the power of an assessment to make a change in society requires an understanding of the feedback loops the consequences of an assessment have on the individuals assessed and the institutions doing the assessing.

Similarly, assessment can be understood as a process of *inquiry*. Assessments can be used to produce and evaluate evidence about the capabilities of students and communities. This metaphor envisions moving from the practice of assessing predefined qualities to the use of assessment to discover and better understand phenomena that affect learning. It proposes a more open evidentiary framework, such as the one used in the AP Studio Art program, to reveal differences between examinees and understand the reasons for them.

Conclusion

Different kinds of assessments are used in different ways and for different purposes. Therefore, discussion of assessment and assessment policy cannot be limited to the surface features of test taking. This paper has aimed to provide metaphors to organize thinking about assessment. Awareness of the interconnected key ideas from those metaphors can make us aware of conceptual frameworks we can take advantage of to reason sensibly about assessments in order to better design them to meet our goals.

Assessment as Evidential Reasoning

Joanna Gorin

Introduction

Regardless of its type, any assessment is a systematic method for drawing conclusions from multiple sources of data or evidence. Assessment in education is often conflated with standardized testing, which assumes that sufficient relevant

information to answer assessment questions can be gathered via decontextualized individual test items. That assumption directly contradicts the American Psychological Association (APA) standard that evidence should be gathered from multiple sources in multiple contexts. The future of educational assessment requires a framework for gathering evidence at multiple opportunities about the variables most likely to impact real-world behavior. This paper aims to advance the concept of evidential reasoning as one way forward.

Assessment as Evidentiary Arguments and Evidential Reasoning

Bob Mislevy introduced evidentiary arguments as a metaphor for assessment in which the quality of an assessment's conclusions is based primary in the strength of the argument the assessment makes based on the evidence it provides. In past decades, educational assessment has been limited to the evidence provided by standardized tests. This paper argues that assessments based on multiple sources of contextually rich evidence, including engagement and sociocultural concerns, will improve our ability to make valid inferences and decisions about student learning and instruction.

Evidential Reasoning

We are understandably uncertain about the abilities of students from different classrooms and backgrounds. The Evidential Reasoning (ER) approach relies on evidentiary arguments and decision theory to analyze arguments made by assessments and improve their design. Six crucial terms structure arguments:

1. *Claim* – the conclusion being argued
2. *Grounds* – supporting evidence for the claim
3. *Warrant* – the chain of reasoning that connects the grounds to the claim
4. *Backing* – support and justification backing the warrant
5. *Rebuttal/Reservation* – exceptions to the claim, description and rebuttal of counter-examples and counter-arguments
6. *Qualification* – specifications of limits or conditionality to the claim, warrant and backing.

The grounds are data collected as evidence to support the claim. The warrant justifies the use of that data by virtue of the backing, which illustrates the data's meaning and utility. Exceptions and limitations of the inference can be described through rebuttals and qualifications. The strength of the inference lies primarily in the amount and quality of the data, and the strength of the backing to support the warrant.

The Assessment Argument

ER can help us fulfill the APA standard for educational assessment's requirement to synthesize data from multiple disparate sources of evidence. The interpretation and use of an assessment is the claim, the data is student behavior, and the warrant and backing are additional information about the item and student, such as existing research about item performance and student cognition. All assessment is a form of evidentiary argument, but the use of ER throughout the design and development process in what Mislevy terms Evidence Centered Design (ECD) improves the assessment argument via a process of designing observational contexts to provide contextualized, high-quality evidence about the student.

ECD makes explicit the framework of assessment components that are implicit to any assessment design, including construct definition, item design and statistical estimation of abilities. The *student model* in ECD defines the claims an assessment aims to make about student ability and its conceptualization of that ability — its structure, development and effect on behavior. The *evidence model* describes the salient features of observable behavior used to update estimates of student model variables. When the observable indicators are highly reflective of the latent variables of interest, the assessment's claim is strengthened. The *task model* defines the characteristics and conditions of assessment tasks. Its goal is to depict and guide the implementation of the environment in which a student will exhibit observable behaviors that correspond closely to what is described in the evidence model. The alignment between the behaviors produced by the task model, the behaviors used by the evidence

model to estimate student characteristics, and the characteristics that inform inferences about ability in the student model determines the strength of the overall assessment argument.

The function of ECD is to make critical assessment aspects more explicit in order to guide the construction of tasks. By breaking an assessment down into parts, more attention is given to some of the assumptions often made when designing assessment, improving the quality of evidence produced by those tasks. This improves the argument an assessment makes about student characteristics, ultimately improving score interpretation and use. Two examples of ECD in modern assessment design follow.

Alternative Assessments

The alternative assessments are given to the 1 percent of students with severe cognitive deficits, and are designed to measure the skills of students who are not expected to perform at grade level. Alternative assessment practices vary greatly, but generally practice better models of evidentiary reasoning than that implemented for the general student population. Alternative assessments draw from behavioral observations, performance assessments, rating scales and portfolios. Each of these indicators requires collection of evidence samples that characterize students' ability. The assessments often consider auxiliary information about the student including past performance, specific abilities and other background information that affects evidence interpretation.

The underlying logic draws on multiple sources of evidence to make claims about a student's proficiency level. Evidence samples are the data. The correspondence between the data and grade-level content standards and the use of multiple raters provides backing and warrants. That said, the assessments have significant weaknesses. The backing for the use of some of the evidence sources is questionable, weakening the argument for the claim. The arguments themselves often are lacking in empirical evidence associating the observed scores explicitly with the targeted skills, as in a student-evidence-task model setup. These problems arise from small, heterogeneous student populations and variable practices.

Efforts are underway to develop measurement models to improve alternative assessment and its ability to measure the growth and status of students with disabilities. The approach of considering broad, contextualized sources of evidence is a base for moving forward. The lessons from alternative assessment are instructive for adapting general education assessments to incorporate multiple sources of evidence.

Psycho-educational Assessment

Psycho-educational assessment (PEA), like that performed by school psychologists and counselors, also offers an example of multi-source evidence-based reasoning. The goal is to diagnose an underlying cause of "atypical" behavior and determine a positive course of action. PEA proceeds using a framework that draws on multiple data sources to make a series of yes-no determinations, the probabilities of which change based on the presented evidence. The goal is to make a probabilistic claim about the most likely diagnosis.

The system uses varied sources of evidence, from third-party observations to scores on standardized tests. Both quantitative and qualitative evidence is included. Except for standardized testing, the evidence is obtained from the student's authentic daily tasks and contexts, which are the ones PEA is making claims about. The process merges measurement science and clinical assessment, with each field offering backing and warrants for distinct sets of evidence that contribute to PEA's argument. The decision-making process by which both kinds of evidence are collected and integrated is critical to arriving at a valid claim.

The goal of PEA is to support decisions regarding students in classrooms. In order to do so, it relies heavily on classroom behavior, distinguishing it from and comparing it to more "context-free" clinical measures. The varied data sources are needed to choose among competing explanations for the student's issue. The use of both standardized measures and classroom-based data is necessary to select an appropriate intervention.

Although educational assessment for accountability typically lacks the diagnostic goal of PEA, the use

of multiple, varied data sources and evidence from appropriate contexts should transfer well to accountability assessment. Every assessment claim is a form of diagnosis. The use of educational assessment to inform instruction must rely on incorporating the learning context into scoring. Otherwise, the data collected can only be used to make claims about abilities in isolation and not in contextual use.

Implications for Assessment Design and Development Practice

The use of evidential arguments for educational assessment has practical implications. The future of educational assessment will be driven by:

- Changes in the nature of claims we make about students
- Availability of new data sources to inform our argument
- New analytic tools to translate data into evidence

New Assessment Claims

A high-quality argument includes only evidence relevant to its claim. Data that is informative for one claim may be irrelevant for another. We have to consider whether the assessments we currently use provide evidence regarding what we want to know about students.

The constructs used in traditional high-stakes standardized assessments focus on a narrow band of "basic skills" and their operational definitions. Regardless of the intent of this focus, its result has been a narrowing of the curriculum and the de-emphasis of elective and higher-order content areas. The ultimate effect has been neither to improve U.S. standing in international tests of those basic skills nor to produce graduates able to deal with employers' demands for problem solving and critical thinking.

Current assessments don't provide sufficient evidence to make strong claims about students' ability with higher-order, 21st-century skills. Unlike most basic skills, the processes encompassed by higher-order skills are multidimensional, cross-contextual and cross-disciplinary. They emphasize the ability to apply knowledge in the service of goals larger than answering a test question. In order to make a meaningful claim about students' higher-order abilities, we require evidence that is useful for describing and predicting students' real-world, contextualized behavior.

Traditionally, our theories about students have focused on individual differences, the acquisition of knowledge in response to instruction and cognitive models of student thinking. Making claims about students' contextualized performance will require a new theoretical focus on the interactions between individual cognition and the situative context. New models of behavior have to consider individual, situative and cultural characteristics simultaneously. This is a broader, more demanding evidentiary requirement than traditional models, and requires the development of new sources of evidence.

New Evidence Sources

Educational assessment has relied on scored responses to paper-and-pencil group-administered tests due to early technology and demands for testing of large numbers of military recruits. Little has changed in the past century: with the exception of adaptive testing and the use of constructed responses, testing remains the paper or computer administration of identical items measuring unidimensional constructs. The approach yields little evidence for claims about an individual's ability to reason in the real world or eventual success in higher education or the workplace. This argues for the need for new kinds of evidence more closely tied to the types of claims employers and policymakers want to make. The use of multiple evidence sources and varied data across multiple contexts can provide more robust arguments about student learning.

Twenty-first-century technology has expanded our capability to capture and analyze data about students and the assessment context. The constructed-response and performance tasks becoming popular today are facilitated by automatic scoring systems. However, they still suffer from a limited set of observations in an artificial context. They lack appropriately situated contexts to generate evidence of how students reason in natural environments.

New assessment tasks should incorporate items that require processing consistent with contemporary

models of student cognition. Among those are scenario-based tasks, simulation-based assessments and educational games. Scenario-based assessment, used widely in the professional disciplines, embeds multiple-choice and constructed-response questions in a realistic context. Simulation-based assessment goes further to mimic a real-world situation as closely as possible, allowing examinees to interact with tasks in a way analogous to the real-world environment. Successful simulation-based assessments like Cisco's Networking Performance Skill System mimic actual equipment to measure network proficiency and provide feedback to students and instructors.

Educational games aren't based in realistic environments, but can still be powerful assessment tools. They offer a structured learning space in which complexity and sequencing of objectives can be controlled. They can increase student interest and engagement. They often play out in a social context, requiring students to learn rules of engagement and community standards. Games can provide valuable evidence of a student's ability to strategically navigate different contexts.

Creating a task environment that includes the skills we're interested in measuring can strengthen the assessment argument. Computer-based assessments can leverage technology to capture more data on student processes and interactions with assessment tasks, such as response times and student log data on keystrokes, mouse clicks and scrolling. That information can be captured and examined as evidence about student learning. In addition to certification and scoring, process data is useful for diagnosing specific problems and student misunderstandings. Establishing the criteria by which by which student logs can be compared is an area that needs further research.

Some research has moved into collecting psycho-sensory data, including pupil dilation, eye movement and brain activity. While still rare, that data can provide valuable evidence of cognitive processes and attentional resources. More research on the backing and warrants for those evidence sources is still needed before that data can be used as part of the assessment argument.

New Analytic Tools

To make claims using evidence, the data must be translated into a useable form. The traditional approach focuses on transforming scored items to estimates of latent traits. Providing useable evidence to support more complex interactionist claims requires multi-dimensional models incorporating multiple pieces of data. Statistical modeling techniques such as Bayesian inference networks provide a probabilistic model for incorporating multiple and heterogeneous evidence into an argument.

Conclusion

Evidentiary arguments are defined by the claims they seek to substantiate. Traditional educational tests make claims based on very narrow evidence about students' ability to correctly answer isolated questions. If the goal of education is broader than that, then we require assessments that parallel the complex cognitive and sociocultural learning we value. The view of assessment as a one-hour, one-day or one-week effort must end. The future of assessment should look more like everyday real-world interactions than our typical notion of an educational test.

Preparing for the Future: What Educational Assessment Must Do

Randy Bennett

Introduction

The mechanisms, delivery and content of education are all changing in response to the rise of digital technology. Assessment must change as well, or education and assessment will increasingly work against one another. This paper advances the claims that assessment must:

1. Provide meaningful information
2. Satisfy multiple purposes

3. Use modern conceptions of competency as a design basis
4. Align test and task designs, scoring and interpretation with those modern conceptions
5. Adopt modern methods for designing and interpreting complex assessments
6. Account for context
7. Design for fairness and accessibility
8. Design for positive impact
9. Design for engagement
10. Incorporate information from multiple sources
11. Respect privacy
12. Gather and share validity evidence
13. Use technology to achieve substantive goals

1. Provide Meaningful Information

In order for teachers to make effective decisions about instruction and for policymakers to make effective decisions about institutions, assessments have to provide accurate, actionable information. Information drawn from assessment has to be based on a strong evidentiary argument in order to support actions that can meet the goals of education. The growth of international comparisons increases the calls for assessments that can inform improvement of the educational system. Formative assessment also is a growing area of interest because of its potential to inform interventions throughout a course, rather than only at the end of it.

2. Satisfy Multiple Purposes

The previous claim's use of "meaningful information" is oversimplified: moving into the future, assessments will have to produce multiple sets of information to fulfill multiple purposes. Education officials need different information for evaluating schools than teachers need for differentiating instruction and improving practice. No one assessment can fulfill those diverse purposes, and one designed for multiple ends may be optimal for none. A set of different, related assessments designed to work together systemically holds more promise for satisfying the different goals of education.

3. Use Modern Conceptions of Competency as a Design Basis

The list of general and domain-specific competencies that society considers important is evolving. Assessments must be designed with our most up-to-date understanding of what it means to be proficient. Too many current tests are based on mid-20th-century ideas of competency that are outdated and undervalue what we now know about how knowledge is constructed, organized and presented in the procedures used to solve problems. Greater attention to new research on learning progressions could increase the relevance of test results for educators. Content standards alone cannot effectively guide test design.

4. Align Test and Task Designs, Scoring and Interpretation with Those Modern Conceptions

Design according to modern conceptions of competency requires the development of competency models, more research on instructional practices that align with those models, and a deliberate effort to link models and practices when designing the components of an assessment system. Traditional item formats will be insufficient for this task. Simulations and other extended constructed-response formats are necessary to observe examinees' capacity to solve problems using multiple competencies simultaneously. Smaller, discrete tasks within a simulation are needed to help students and teachers discover which sub-competencies might be responsible for success or failure on a complex task. Future scoring mechanisms will need to be designed specifically to recover evidence from performance on those tasks.

5. Adopt Modern Methods for Designing and Interpreting Complex Assessments

Methods like Evidence-Centered Design and Assessment Engineering are examples of modern design methods that offer evidence-based models for thinking about assessments, examinee performance and using them to produce useful data. They begin with specifying the claims an assessment aims to make about individuals or institutions. Next they detail the evidence needed to support those claims, and then finally the tasks required to elicit that evidence. The accumulation of evidence from specific tasks about specific examinees gradually provides the basis for the claim the assessment ultimately makes about that examinee or their school.

The use of such a model requires school systems to think about assessments that, rather than rank students along a single dimension, provide multiple strands of information at once. Rather than being standardized, such assessments can be dynamically assembled from a data library of task models, rubrics, and evidence models. This provides for more flexible delivery of authentic tasks than traditional assessment and will require deep changes in the traditional testing and scoring architecture used by schools.

6. Account for Context

Student performance on an assessment can be measured, but interpreting why a student performed as he or she did requires an understanding of the student's context. Making comparative decisions about school accountability and college admissions using standardized tests in standardized conditions has resulted in assessments far removed from the instruction students actually encounter. Factoring students' GPA and personal statements into qualitative judgments by admissions professionals is one way to address the issue, but federal state and accountability efforts typically offer little insight into the instructional or home context of students.

Embedding assessment into the instructional context should produce more actionable information for educators. However, formative assessment data may not be as useful for aggregated school or district performance reports. It also is ill-suited for consequences-based accountability efforts, and may pose privacy issues related to ongoing surveillance of student performance.

7. Design for Fairness and Accessibility

Standardization is meant to provide fairness in evaluation and comparisons. Early standardized tests in the United States didn't value fairness between groups — and indeed were often used for racist purposes — and with the exception of the SAT, the measurement community didn't make addressing group-level fairness a priority until the second half of the 20th century. Concern for minority groups, students with disabilities and English language learners will factor into the design and use of both formative and summative assessments. Indiscriminate use of differentially valid assessments can increase achievement gaps if the interventions they inform differentially impact different groups. Government and advocacy organizations must address instances of differential validity that disfavor underserved groups.

8. Design for Positive Impact

Tests have profound implications on the behavior of students, teachers, and organizations. The NCLB Act was premised on an intended positive impact. However, tests can have unintended effects: NCLB is commonly criticized for curricular narrowing caused by the interaction of the law's focus on reading and mathematics, different state content standards, limited methods to measure achievement and the threat of sanctions.

If higher-quality content and assessments are to have a positive impact, then summative and formative assessments have to be built that model good instructional practices. That requires:

- substantive, realistic tasks
- the inclusion of tools and representations routinely used by proficient performers
- tasks that explicitly connect qualitative and formal understandings
- test structure that demonstrates the scaffolding of complex performances

- the use of learning progressions to measure changes in student understanding

Tests that faithfully recreate the situations and competencies education aims to target can be beneficial even if students are "taught to" the test.

9. Design for Engagement

Assessment results are more meaningful if students give maximum effort. Designers can drive student engagement through the use of problems students care about, motivational feedback, multimedia and game elements, and hardware with which students are already well acquainted. Games designed to generate meaningful evidence for adjusting instruction or measuring competencies can be especially useful. As seen in the AP Studio Art exam, common scoring frameworks can be used to aggregate information across students playing games that are different on the surface but that are designed to measure the same skills.

10. Incorporate Information from Multiple Sources

All assessments take a limited sample of the full range of student behavior, and each assessment method is subject to its own specific limitations. As such, multiple sources of information are needed to generate meaningful information that can be used to make consequential decisions, as is already the practice in college admissions. This is true not just of summative assessments, but for formative assessments as well. Rather than adjust instruction based on a single interaction, a teacher should form a "formative hypothesis" that is confirmed or refuted through other observations of classroom behavior, homework, quizzes or other tasks. Technology will increase the kinds of information available to teachers for these judgments.

11. Respect Privacy

Some commentators laud the possibility that technology will enable ubiquitous, surreptitious assessment that will eliminate the need for disruptive stand-alone tests. Despite the idea's attractiveness, test designers must keep in mind the privacy rights of students and teachers. Individuals should know when and why they are being assessed. Without a space free from assessment, students and teachers may feel their ability to experiment in teaching and learning has been constrained. Whether learning behavior is within the power and rights of the state to observe and use for consequential purposes is an open legal question. There is a distinction made in sports between performances that count towards player statistics and team standing and those that do not, and a similar compromise may be appropriate for education.

12. Gather and Share Validity Evidence

Whatever their strengths, future assessments will need to provide evidence to support their results in order to gain legitimacy in the educational community. Independent ongoing analysis of the meaning, impact and quality of assessment results is necessary in order to gain the trust of all of the stakeholders involved. Score-generation methods cannot be held so closely by test vendors as to prevent independent review by measurement experts. Only transparency will allow those experts to report to the wider education community about the quality of the tests.

13. Use Technology to Achieve Substantive Goals

Technology in assessment is useful only so far as it enables us to accomplish what couldn't be accomplished with traditional tests. Measuring new competencies and evaluating existing competencies more effectively are among those goals. The use of technology without a specific plan to positively impact teaching and learning should be avoided.

Conclusion

Education is changing quickly, and assessment must keep pace or become irrelevant. Adapting and reinventing our assessment systems is necessary for education to achieve the goals we've set out for it as a society. The challenge to the measurement community is to retain its foundational principles while applying them in new ways to meet the decision-making requirements of a dynamically changing educational world.

The Gordon Commission
on the Future of Assessment in Education

Contact Us

Please send us your feedback,
comments and suggestions.

Email: contact@gordoncommission.org

Gordon Commission
P.O. Box 6005
Princeton, NJ 08541

The Gordon Commission was established by ETS to investigate and advise on the nature and use of educational testing in the 21st century. 21186