The Gordon Commission
on the Future of Assessment in Education

Four Metaphors We Need to Understand Assessment[1]

Robert J. Mislevy
Educational Testing Service

[1] Prepared for the Gordon Commission on the Future of Assessment in K-12 Education. I am grateful to John Behrens, Randy Bennett, Jim Pellegrino, and Lorrie Shepard for comments on an earlier version.

1. Introduction

Everyone agrees that assessment is integral to education, but discussion about how to improve it is prone to contention and people talking past one another. A major problem is that people don't share a language that is rich enough to discuss interpenetrating issues that arise at many levels—personal, social, educative, economic, technological, and statistical, to name a few. People who may be quite knowledgeable in their own areas of expertise draw mainly on their personal experience as students with familiar kinds of assessments (Braun & Mislevy, 2005). This intuitive frame suits familiar contexts, but it falls short for discussing assessment policy or for thinking about different kinds and contexts for future assessments.

My goal in this paper is to provide a quick-start guide to four metaphors[2] for understanding assessment. Each offers a set of concepts, relationships, processes, and actors for thinking about real-world situations. Applying their perspectives to assessment systems brings out assumptions and implications we might have otherwise missed. The four metaphors are these:

Assessment as Practice

Assessment as Feedback Loop

Assessment as Evidentiary Argument

Assessment as Measurement

I suggest these metaphors after a career of opportunities to learn about assessment the hard way—hundreds of assessment applications, some successful and others not, from formative mixed-number subtraction tests and the National Assessment of Educational Progress, to Advanced Placement Studio Art portfolio assessment and an intelligent tutor for troubleshooting the F-15 aircraft. Some brought out challenges at the statistical, psychological, social, and epistemological foundations of assessment. The metaphors have helped me design and use assessments, and they are useful for thinking about assessment policy and assessment futures. I will draw illustrations from projects I've worked on, not because they are always the best examples but because I think I understand them.

The next section summarizes all four metaphors. Sections 3 through 6 discuss each in more detail, noting their interconnections and their relevance to issues in assessment reform.

[2] Calling them semantic frames (Fillmore, 1976) would be more accurate but less welcoming.

Section 7 points the insatiated reader h to four more metaphors, less central but also useful for thinking about the nature and the practice of assessment.

2. Short Descriptions of the Four Metaphors

2.1 Assessment as Practice.

The term *practice* comes from a sociocultural perspective in psychology. Practices are recurring, organized activities that people become attuned to, learn their constraints and affordances, and use to interact with other people and situations. Two implications of seeing assessments as practices are critical to any discussion of assessment design and assessment policy. First, as a practice itself, the capabilities that an assessment requires from students will share some of the capabilities of the real-world practices it is meant to prepare them for—but it is not the same as those practices. What people learn can be surprisingly bound to the conditions and the practices in which we learn it. There are students who can use algebra in algebra tests but nowhere else in the world. Second, as a practice occurring in the social world, an assessment sets expectations, influences learning, channels resources, and shapes the very way people think about students and capabilities. This contextualization shapes the meanings of all the variables and inferences in any assessment system, from informal one-on-one coaching sessions, to standardized tests, to investigations in simulated worlds. "Assessment as practice" both complements and counters "assessment as measurement."

2.2 Assessment as Feedback Loop.

This metaphor makes us see assessment from the perspectives of people in different roles in an assessment system, who need to use information about students' learning. This can be learners themselves, in learning games or tutoring systems. It can be teachers using formative assessments in their classes, or chief state school officers using standardized accountability tests for funding decisions. In each case an actor uses information at some timescale, at some grainsize, and in some context: faster, more detailed, and more deeply contextualized for a tutor working with a student; slower, coarser, and with little context for the chief state school officer's annual survey. The essential questions are who needs what information, when, for what reasons, and what other information do they have to work with? Different answers to these questions can

lead to different forms of assessment, with different properties, different purposes, and different expectations. A key insight from this metaphor is that the value of assessment data is not inherent in either the assessment or the data themselves. It depends on who's using it for what decisions, in light of the other information they have about the learner and the context. The same information can be valuable feedback to a person with one role in a system and worthless to someone in a different role.

2.3 Assessment as Evidentiary Argument.

The concepts and machinery of evidentiary reasoning help us understand assessment in context. We can adapt concepts from philosophers such as Stephen Toulmin and evidence scholars such as John Henry Wigmore and David Schum to assessments, both familiar ones and new ones (Mislevy, Steinberg, & Almond, 2003). Evidentiary reasoning provides a coherent framework for reasoning from the particular things students say, do, or make in a limited set of situations, to what they know or can do more broadly or what they should work on next. The argument metaphor for assessment first appeared in Lee Cronbach's and Sam Messick's writing on validity. Current work is proving helpful to develop new forms of assessment such as simulations and games. This metaphor connects the situated and practical perspectives of the practice and feedback metaphors with the engineering toolkit of the measurement metaphor.

2.4 Assessment as Measurement.

Measurement has been coupled with assessment for over a century. In the beginning, measurement was taken literally. Mathematical models were developed to evaluate test items, to gauge the accuracy, to construct tests, and to reduce biases in test design and test use—all presuming that well-defined, existing, quantitative properties across students were there to be measured. Developments in psychology challenge this presumption. Educational measurement is better understood not as literally measuring existing traits, but as providing a framework to reason about patterns of information in context. These patterns emerge from the dynamic interactions among students and situations. Appropriately interpreted, measurement models can nevertheless guide assessment design and assessment use in technical ways that the other metaphors cannot. The measurement metaphor provides engineering tools to manage data and

uncertainty, to support gathering and reasoning from complex assessments and extensive data, and to pursue fair and effective use of assessments no matter what form the assessments take.

3. Assessments as Practices

Assessments are organized social activities that are integral to learning and decision-making situations. The "assessments as practices" metaphor builds on what we are coming to understand about how people learn, think, and act, and how they interact in the physical and social world. A coherent view is emerging from such diverse fields as linguistics, situative psychology, cognitive anthropology, comprehension research, activity theory, and neuroscience. This section offers a brief sketch of a psychological perspective that underlies the "assessments as practices" metaphor, then looks at some implications for thinking about assessment.

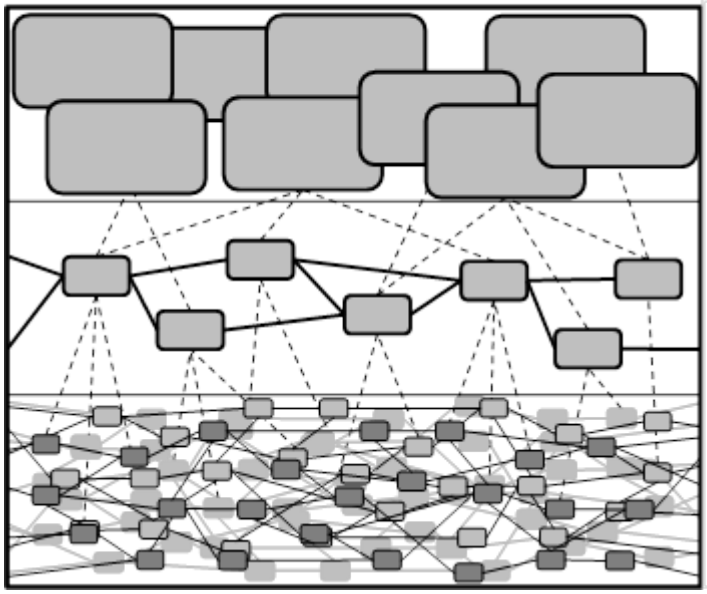
3.1 The sociocognitive perspective[3]

Figure 1 depicts three levels of phenomena and associated time scales (Lemke, 2000). The middle layer represents the actions, events, and activities we experience as individuals—that is, human-level activity, people acting within situations. We interact with the world and with each other: thinking, planning, conversing, reading, working, playing, solving problems, using representations, and cooperating or competing with family, friends, co-workers, and others we do not know personally. Assessments activities take place at this level.

All of this activity is mediated by the between-persons patterns suggested in the top panel. These are regularities in the interactions of people in their overlapping identities and communities, and it is through them that actions constitute meaningful activities (Wertsch, 1998). There are widely shared conceptions such as what it means in a culture to be sick or to be married, and more focused patterns of activity in classrooms and grocery stores. There are narrative structures, from common themes in human interactions to highly structured scientific models. There are tools, languages, and other semiotic systems. There are fine-grained patterns such as arithmetic schemas, and the grammars and constructions. I will use the broad term "linguistic, cultural, and substantive" (LCS) patterns to encompass these ways of thinking and

3 The term "sociocognitive" sounds a lot like the more frequent term "sociocultural" as a modifier of psychology. "Sociocognitive" focuses on the interplay between what is happening within persons, cognitively, and what is happening between persons, socially. Greeno's (1998) discussion of a situative psychological perspective is what I have in mind. When writers such as Gipps (1999), McNamara and Roever (2006), and Foucault (1977) talk about sociocultural aspects of assessment, they address the substance and the processes of the interplay of the social and cultural, rather than the processes in the interplay between individual cognition and culture. This paper offers brief comments in Section 7 with regard to Foucault's "assessment as the exercise of power" metaphor.

acting. Communities and disciplines are marked by identifiable, recurring, clusters of themes, structures, and activities, called *practices*, from brushing teeth to writing grant proposals – and building, taking, and using assessments.



Time Scale	Interaction	Examples
Millenia Centuries Years Months Weeks Days Hours Minutes Seconds	Societal change: Language, culture, and beliefs change over time as individuals and groups interact through them. Human-scale activity: Individuals act and interact, think, etc. in socio-culturally framed situations	Evolving language; cultural norms; social institutions; technological change; scientific models Conversation, problem-solving, business transactions, driving, studying, assessment
Milli-seconds	Neural activity: Within-individual events and processes that produce cognition	Perception, comprehension, learning, feeling

Figure 1. Levels and time scales in human activity.[4]

The bottom layer represents the within-person processes that produce individuals' actions. To produce successful human-level activity, a person's neural patterns must be able to relate to both LCS patterns and adapt to suit unique situations. Young (2009) uses the term *resources* to refer to a person's capabilities to assemble patterns to understand, create, and act in particular kinds of situations. A person develops resources by participating in practices (Lave & Wenger, 1991). In this way people become attuned to LCS patterns, their affordances and constraints, when, how, and why people use them, and how to use them themselves.

4 Why is problem solving shown at the middle layer and learning and comprehension at the bottom? I want to reflect the distinction that Stanovich and West (2000) call System 1 and System 2, Norman (1993) called reflective and experiential cognition, Kahneman (2011) calls slow and fast thinking. I am referring here to problem solving as thinking consciously through steps of a problem, reading instructions, fitting models, etc., at a time scale between minutes and weeks. I am referring to comprehension and learning in terms of the neural processes of activation and integration in, for example, Kintsch's (1998) model of comprehension. It is certainly true that some problem solving is experiential, and slow problem-solving consists of coordinated long sequences of fast thinking; and reflective thinking guides our attention and action as we work on comprehension and learning over time. There is no clean break, but a useful point is made with this placing of the terms on the scale.

Instructional activities, assessment activities, and real-world activities all take place in the middle layer.[5] They are distinct, yet intimately related, practices. Despite having distinct contexts and standards, an instructional practice shares some key LCS patterns and activities with real-world situations. It is thus meant to help students develop resources they will be able to draw upon in real-world situations. Assessment is observing students in yet another set of situations with their own contexts, standards, and activity structures, again sharing some LCS patterns with both the instructional and real-world situations. The intent is to observe students acting in a handful of assessment situations, interpret the actions through the lens of the targeted practices, and make inferences about the person's capabilities to act in relevant real-world situations or learn in instructional ones. The myriad ways of designing, administering, taking, and interpreting assessments are all particular kinds of practices. We can think about any particular assessment in terms of the interplay among targeted interpersonal LCS patterns and students' personal resources that underlie their activity. We can then think about how well the activities support the intended purpose of an assessment.

3.2 Some implications for assessment

The capabilities an assessment requires will share some of the capabilities of the real-world practices it is meant to prepare students for, but the assessment itself is not the same as those practices. What we learn can be surprisingly bound to the conditions and the practices in which students learn it and assessments test it. Some students can use algebra for algebra tests but nowhere else in the world, and others can use it in certain real-world situations but not in tests (e.g., Saxe's (1988) candy-sellers). An assessment signals to students what is important to learn and what the standards of good work are—for the practice of that assessment. It is an empirical question whether this is the same as learning the standards of good work in the real-world practices the assessment is meant to relate to.

This means we cannot assume that students will develop the kind of thinking and acting we care if we do not engage those capabilities in assessment and instruction. Conversely, instruction and assessment that do engage students in the targeted thinking and acting are more likely to develop the targeted capabilities. The more assessment situations differ from the real-

[5] And they are often intertwined, not so separate as the point being made might suggest. A one-on-one tutoring session is both instruction and assessment—seamless, informal, and individualized. Dissertation research in the biochemistry lab is also instruction and assessment as well, and valuable practical work to boot.

world situations, even if the shared elements match closely, the less we can count on transfer and the weaker the inference from assessment performance will be. This is the argument for performance assessments, contextualized assessments, assessments based on extended work and self-directed work. This is why David Williamson Shaffer (2006) designs his "epistemic game" simulations to help students develop not just the knowledge and skills in a domain like journalism or urban planning, but also the identities, values, and epistemology of the community.

It is harder to assess identities, values, and epistemology than it is to assess knowledge and skills, and for the same reason it is hard to assess so-called higher-order or 21st Century skills such as communication, problem-solving, and information literacy more generally. They are not well-defined skills that everyone will develop in the same way and can demonstrate with the same performance to the same task. Rather, they are qualities of thinking and acting that are manifest in different ways in different domains and different situations. Indicators might include instances of, for example, entertaining multiple perspectives, proposing and justifying choices, or anticipating and countering objections to a course of action[6] – each interpreted in light of a situation and an individual's history relative to that situation. We can anticipate that the strongest evidence will be most contextualized, and the weakest what can be gleaned from external assessments that are not connected with students' instructional contexts or histories; and concomitantly, the stronger the evidence the less portable it will be. Indeed, rather than trying to assess such capabilities per se, an alternative is to focus assessment on knowledge and skills (which includes knowledge representation, reasoning from evidence, etc.) but to do so in a way that is consistent with real world practices and developing identities.

Consider for example "intellective competence" as defined by Gordon (2008, p. 21): "I have come to use the term to refer to a characteristic way of adapting, appreciating, knowing, and understanding the phenomena of human experience. … These developed abilities are reflected in the student's ability and disposition to use knowledge, technique, and values through mental processes to engage and solve both common and novel problems." We should not expect a decontextualized standardized test of intellective competence. Any assessment of intellective competence will by necessity be a contextualized application of the abstract qualities in the definition. As Section 4 notes, this approach suits some purposes and contexts of assessment but not others.

[6] Personal communication, E.W. Gordon, January 17, 2012.

Assessments are components of sociocultural systems. This fact connects with the feedback metaphor in obvious ways for the effects that people have in mind for assessments. But assessment also has powerful effects on people and institutions in the subtle ways it influences other practices. Assessments influence how people think about the nature of learning, how they study, what they think is important in a domain – and what happens day by day in the classroom, and second by second in a student's brain. The "assessment as practice" metaphor makes us aware that assessments do not simply measure existing qualities in students, and they don't even just shape the development of those qualities. Rather, in a degree that is arguable but an effect that is not, they *cause* those qualities to exist, and peoples' lives and practices to adapt to them. This observation underlies the policy strategy of using assessment as a lever for reform that Section 7.3 touches on.

4. Assessments as Feedback Loops

4.1 The basic idea

Education is about improving the capabilities of students. Doing this effectively requires a variety of decisions, at many levels and by different actors: Teachers in classrooms, to be sure, and students themselves; but also higher-level educators such as principals and chief state school officers; so too policy makers, employers, admissions officers, and the public at large. Each in their own way need information about how educative efforts are faring, in order to evaluate them, allocate resources, or decide what to do next.

Any assessment is meant to gather information for some actor, for some purpose, under some constraints, and with some resources. Assessments can differ markedly in these ways. Exactly the same assessment can be invaluable for one place and purpose, but worthless for another. To design or to evaluate an assessment, then, is not simply a matter of looking at tasks but considering how to best provide information to whoever needs it, in light of what they know and what they need to do. The evidentiary argument and measurement metaphors help tune an assessment to its purpose, but one can neither evaluate nor design an assessment without understanding its intended role in some feedback cycle.

4.2 Examples

A one-on-one tutoring session is highly contextualized and highly individualized. The tutor at all times considers "What can this student be thinking so that what he just did makes sense?" The tutor may ask a question, suggest an action, explain a concept, rephrase a student's comment, or encourage a next step with a silent nod, all to the end of providing an experience that extends the student's resources. This in-the-moment assessment can be tuned precisely to the immediate situation, the history of the student, and the decision options that are available. But a video clip of the same small interchange that constitutes a breakthrough in the session may be unintelligible to an outside observer. Ten thousand clips, of different students working on different issues on different days under different circumstances, would overwhelm an external observer, and thus be inadequate as a summary assessment of learning.

A classroom quiz is designed to provide feedback to a teacher at the level of the class, to shape decisions about subsequent instruction for the group and sometimes individual students. It is less contextualized than a tutoring session. It does take into account the teacher's instructional options and what students have been working on, and it does leverage their understanding of the expectations and standards of the practice. On the other hand, the tasks are the same for all the students in the class and they are all responding at the same time. This requires less of the teacher's time and makes summarizing information across students easier, but it provides less focused information for each individual. It is an assessment form with a different signature of resource demands, contextualization required, information provided, and decisions served than a tutoring session. It is a practice that has evolved over the past hundred years that proves effective for a particular situation that occurs in the practice of classroom instruction as it happens to exist in most schools today.

The classroom-quiz example introduces the notion of standardization. In common parlance, "standardized testing" refers to assessment bearing a particular constellation of features: High-stakes uses, externally imposed, not contextualized, decontextualized tasks, often multiple-choice formats, time-limited, and sequestered performances. But standardization is not an all-or-nothing characteristic of an assessment. Every assessment has many features, and it may have some subset of features that are the same across examinees and others that are not. Section 5 discusses how these design decisions are made in light of their effect on assessment arguments, and how they depend on assessment purposes and users' existing states of

information. We see in this section how these considerations are integral to the feedback loop(s) an assessment is meant to inform. Standardization is best understood in terms of aspects of assessment structures or procedures that are to some degree determined in advance, because of their implications for assessment arguments. These aspects can concern settings, standards, rubrics, representations, instructions, contexts, and forms of response. (See the box for more examples.)

The SAT is a stereotypical example of a standardized test. It provides college admissions officers with some information about students' reasoning with verbal and quantitative information in a specialized, standardized, setting. The information is much sparser than individualized, contextualized, and extended-over-time body of evidence that go into students' grade point averages. Colleges do not want SAT scores primarily because of what they assesses, even though what they do assesses is surely related to what students will do there. Rather, they want SAT scores because they know the provenance of the information, despite the limitations of the information itself (which is why SAT cheating reports are so damaging).

Examples that Illustrate Points about Standardization

The ASSISTment intelligent tutoring systems (www.assistment.org; Razzaq et al., 2005; Razzaq et al., 2007) offers instructional support to students by introducing a set of scaffolding questions and making available informative hint messages as students work on assessment tasks. Some of the interactive tasks that ASSISTment presents look on the surface like tasks that an advanced computer-based standardized achievement test in, say, algebra might present. The feedback that students get for wrong or incomplete answers is, like the tasks, predetermined. ASSISTments are an economical-to-build learning/assessment because their content and their tailored interactions with students are predetermined, but powerful because the feedback loop they serve with these tasks is tight and contextualized, determined minute-by-minute to what a student is working on and what might be done next. Although the *substance* is standardized, technology is capitalized on to make the *timing* for each student optimal for that individual. It is not so much the content of the tasks that provide the value, but their use at just the right time and the right place for each student in an individualized feedback loop.

The National Board of Medical Examiners (NBME) uses two innovative assessments in the United States Medical Licensure Examination (USMLE) system, each standardized in its own ways. In the first half of the 20th Century, medical licensure in the United States employed a "practical" examination: A candidate would examine a real patient in a real hospital or clinic, and be evaluated by a real practicing physician. What could possibly be more valid? Yet a generalizability study showed that scores depended more on the patients that examinees happened to be assigned and the physicians that happened to score them than on their underlying proficiencies (Melnick, 1996). For the next forty years, the USMLE relied on standardized multiple choice tests alone, while NBME researched methods by which they could validly and fairly assess the important qualities that the clinical exam was meant to reveal, such as managing patients interactively over time and carrying out procedures and communicating face-to-face with patients.

After decades of research by NBME and scores of other medical education researchers and institutions, two new assessments are now part of the licensure sequence. The first is Primum® computer-simulated patient management problems (Dillon & Clauser, 2009) provide open-ended problems with chronic and acute conditions, take about half-an-hour each of examinee time to complete, and can represent anywhere from half-an-hour to six months of time in the simulation. Automated scoring routines provide levels of reliability that rival those of multiple-choice exams. The second new assessment is Standardized Patient Examinations (SPEs; Boulet, Smee, Dillon, & Gimpel, 2009). SPEs involve an examinee interacting with a live person portraying a case. The etiology, circumstances, and persona of the case are predetermined, and every actor portraying a patient is trained on that case. The interactions unfold uniquely for every examinee, much as they would occur in an actual examining room. The checklist of what the examinee did and did not do, and where appropriate, at what quality, are standardized as well. Each actor is trained in the use of the checklist. This assessment is standardized in many respects, but the actual performance is unconstrained and closely replicates the real-world situations it is meant to provide evidence for.

The Cisco Networking Academy (CNA) developed a software environment called Packet Tracer for students and teachers to construct, configure, troubleshoot, and share computer network simulations (Frezzo, Behrens, & Mislevy, 2009). It is the foundation for learning projects, assessment tasks, and interactive games in the community of CNA classrooms throughout the world. Cisco does not operate individual academies but provides central support over the internet, including instructional and assessment materials. The motivation for Packet Tracer was that prior to 2000, hands-on learning with actual networking equipment (routers, switches, PCs, etc.) varied widely in availability and quality across the academies, especially in underdeveloped areas (Behrens, Mislevy, Bauer, Williamson, & Levy, l., 2004). By providing Packet Tracer-based exercises and assessments over the internet, Cisco provided all students an unlimited opportunity to engage in the interactive experiences central to learning how to think and act like networking engineers. Like NBME's Primum assessments, Packet Tracer tasks are open-ended in operation but standardized with respect to interfaces, standards, and evaluation methods. Like ASSISTments, they can be engaged by students individually, at times and places in their learning that the instructors or the students themselves determine. In this way, CNA has leveraged technology and elements of standardization to create assessments that advance excellence and equity at the same time.

The Advanced Placement Studio Art portfolio assessment supports feedback loops at two different levels in the educational system. It uses a blend of standardized and non-standardized features to support uses similar to both one-on-one tutoring and the SAT in different ways. Not surprisingly, there are compromises for both uses. The AP Studio Art supports both situated classroom practice and large-scale, high-stakes assessment in the following way: The work that is judged centrally at the end of the school year is produced in each of hundreds of participating schools throughout the year, as students and teachers create, discuss, and critique their pieces in the individualized and interactive settings of the art classrooms of the many schools. In addition to knowledge of color and design and skills with materials and styles, the program aims to develop what might be called "intellective competence for artists." In the Concentration section of the portfolio, the student must define her own challenge, explain it and discuss resources she drew on, and investigate the problem in a series of works through the year. Although the experience is situated and used in learning day by day, compromises are felt at the local level through the program's requirement to produce certain numbers and kinds of pieces. And the standards by which portfolios are judged centrally must be reflected in the one-on-one discussions. These aspects of standardization make central rating possible. The compromise felt at the system level is that the reliability coefficients for portfolio scores, based on hundreds of hours of work, are a bit lower than those obtained from a couple hours of SAT testing. The assessment design thus balances support for students' learning, on-site contextualized assessment for student- and classroom-level feedback, and provides summary evaluations that help college personal evaluating the students' accomplishments (Mislevy, 2008a).

James Gee poses a question about assessment and proposes an answer:

> So let's say a kid plays [the video game] Halo on hard. And you know, he plays 30-40 hours and he finishes Halo. Would you be tempted to give him a Halo-test? No, not at all. You'd say that the game already tested him. So let's think: Why is it that we are not tempted to give him a Halo-test, but we are tempted to give that algebra test and use that as the judgment? Well, it is because you actually trust the design and learning of Halo better than you trust the design and learning of that algebra class.[7]

What do we see when we look at Gee's examples through the feedback-loop metaphor? First, within Halo play, deeply contextualized assessment is going on moment by moment. The

[7] Downloaded January 17, 2012, from http://vimeo.com/15732568.

game analyzes a player's moves and provides him feedback in the form of the consequences of his actions as he learns to navigate the challenges in the Halo environment. The game obtains information as feedback for its own purposes as well, to adjust the challenges, respond to his actions, and monitor his progress up the levels. To an external observer who is familiar with Halo, knowing the level he is currently on communicates what he has accomplished so far, and knowing he has completed it says much about what he has learned. As Gee points out, knowing the structure of Halo is integral to these inferences.

"Completing an algebra course" is not a well defined term, however, and to an outside observer this datum alone says very little about the resources the student has developed. There are many experiences called algebra courses, and they differ in content and activities they take up and the kinds of situations students develop resources for acting in. A particular student's accomplishments may be considerable, but the outside observer cannot know this. The student's experiences may have been structured around assessments that provided feedback to the student at some points and feedback to the teacher at others. The course may have been something like Carnegie Learning's Algebra I computer-based adaptive tutor[8], which is structured as tightly as Halo and provides moment by moment feedback to the learner like Halo and periodic feedback to the teacher. When would we be tempted to administer an algebra test, then? When, as an outside observer, we need to compensate for lack of knowledge about a student's algebra experience, or when we need to combine or compare information across programs for feedback at a higher level in an educational system.

One unproductive kind of cross-talk about assessment is suggesting that a system replace one kind of assessment with another, when the two assessments are designed for feedback loops that support different users, inform different purposes, function at different levels in a system, or assume different contextual information. A simple example would be saying instead of the current hundreds of thousands of college admissions tests, students should learn in games with built-in assessments. Now it may be a good idea to revamp the college admissions process, and game-based assessments might play some role in a new system, but good policy cannot result from deciding this without recognizing the feedback loops that are being served and considering how alternatives would replace their function or obviate it. A categorical error produces this cross-talk: Different practices with different roles in different systems are all called the same

[8] Downloaded March 21, 2012, from http://www.carnegielearning.com/secondary-solutions/adaptive-math/

word, assessment. The assessment as feedback metaphor helps us make distinctions that the analysis requires.

5. Assessments as Evidentiary Arguments

5.1 The assessment argument schema

The "assessment as argument" metaphor connects a sociocognitive understanding of assessment as practices situated in social systems with the symbol-system toolkit of measurement (Section 6). It provides a rich framework for understanding how the abstract, decontextualized machinery of measurement models acquires situated meaning in context, and how to use it, critique it, and look for instances where the models are inadequate, misleading, or unfair. The reader who follows up with Schum's (1994) *The evidential foundations of probabilistic reasoning,* or better yet his hard-to-get 1987 *Evidence and inference for the intelligence analyst,* will find a way of thinking about assessment that is rich enough to both understand familiar assessments more deeply and to design new assessments around advances in technology and learning science (Mislevy 1994). The roles of warrants, alternative explanations, and the user's knowledge about the relationship between a student and the assessment situation prove particularly useful.

Messick (1994) succinctly describes the structure of an assessment argument:

> A construct-centered approach would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 16)

Assessment arguments build on Toulmin's (1958) general schema for arguments, shown in Figure 2. A claim (C) is a proposition we wish to support with data (D). The arrow represents inference, where the warrant (W) is a generalization that justifies the inference from the particular data to the particular claim. Note that the warrant runs from the general to expectations for specifics; it is prior to the instance of reasoning from particular evidence. Theory and

experience provide backing (B) for the warrant. We usually need to qualify our conclusions due to alternative explanations (A) for the data, which may have rebuttal evidence (R) that tend to support or refute them.

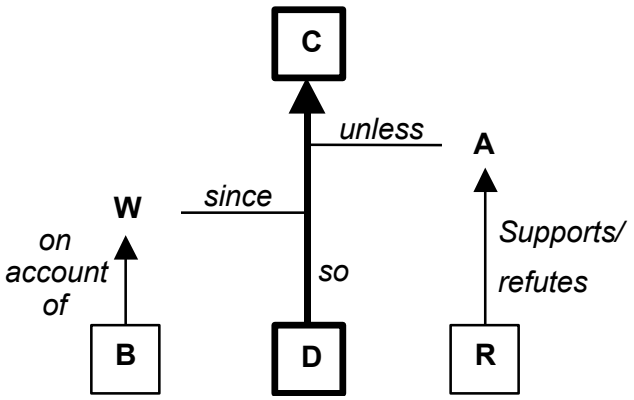


Figure 2. Toulmin's (1958) structure for arguments

Figure 3 applies Toulmin's schema to assessment arguments. At the base of the argument is situated action: an examinee's actions in the task situation. At the top is a claim, based on assessment data, justified through a warrant. The claim is a statement about the capabilities of the examinee. Both the warrant and the claim, as well as features of the situation and the performance, are viewed through some conception of capabilities (Mislevy, 2003, 2006). The warrants are cast in the direction of deductive reasoning: "If a student is characterized as such-and-such with respect to a (possibly multifaceted) construct, then the typical behavior and range of variation of certain aspects of performance in situations with certain features is such-and-such." The flow of reasoning from observed performances is back up through the warrant inductively: "We have observed a collection of performances with such and such features in these particular situations with these features, so our beliefs about the student's capabilities in terms of this construct are such-and-such."

Three kinds of data ground the claim in the assessment argument: (1) aspects of the examinee's actions in the assessment situation, which may include products and processes, (2) aspects of the task situation as seen from the assessor's perspective, and (3) additional information the assessor may have about the examinee's history or relationship to the situation. There can be multiple pieces of each kind of data. Figure 4, for example, shows the multiple and serially-dependent features of performance and situations in extended tasks like role-playing

conversations in language testing and troubleshooting problems in simulation-based assessments. Some key features of a task are designed in by the task developer, but others (both from the perspective of the assessor and the student) come into being only as the event unfolds and may therefore be different for different students.

The last kind of data, "other information," does not show up directly in the formal elements measurement models. Nevertheless it always plays a crucial, if hidden, role in assessment reasoning. We saw in Section 4 that the degree of contextualization and the standpoint of the user are integral to understanding how an assessment functions to provide feedback. It is at this juncture in the argument that the same assessment, the same performance, and the same scores can provide different information to users with different standpoints of knowledge, or who have different uses in mind. Figure 3 shows dashed lines connecting additional information to the interpretations of performances and situations. The dashes suggest that it is a design decision to use or not use this information, and if so what and to what extent. Not using it corresponds to decontextualized reasoning for these links in the chain of reasoning. Even in this case, determining the targeted testing population constitutes a degree of contextualization.

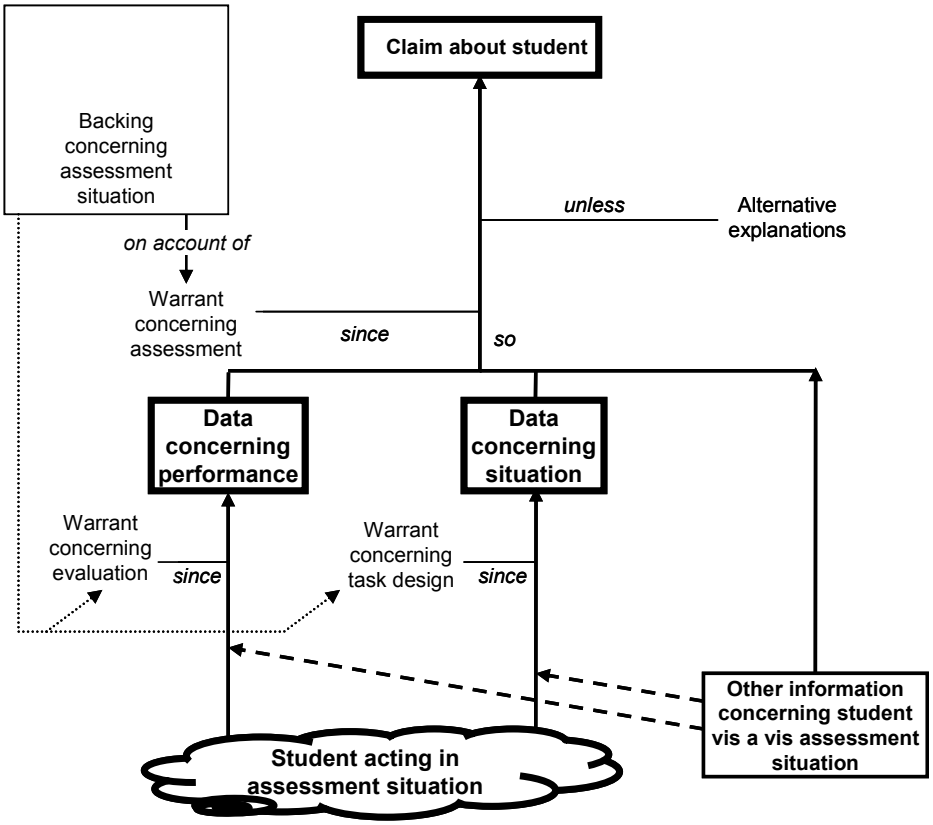


Figure 3. An assessment design argument.

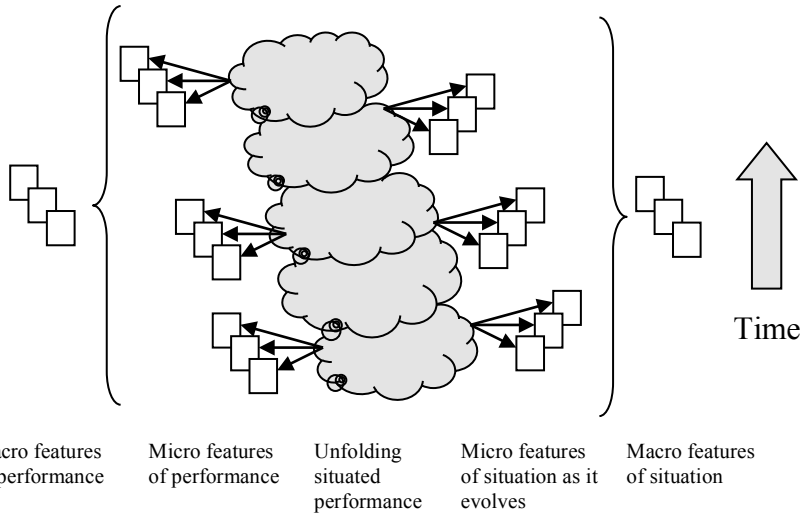


Figure 4. Detail of features of performance and situation in a task that is interactive and/or evolves over time.

Additional information about a student's previous experiences can condition an observer's understanding of his performance. Using an algebraic formula to add the numbers from 1 to 100 is evidence of rote application from a middle school student who has just studied that lesson, but evidence of intellective competence from young Carl Friedrich Gauss who invented it on the spot in grammar school. Poor performance in a complex task can have many explanations that weaken an outside observer's inference, but which a student's teacher is able to rule out. This is a place in the argument schema where we can express an issue that arises in the feedback-loop metaphor: Different observers (e.g., the examinee herself, her parent, her teacher, and the chief state school officer) have different knowledge about the examinee and the situation, and this information affects the nature and the strength of inferences they can draw from a performance.

Warrants are the glue that holds evidentiary arguments together (Schum, 1994). They are the vehicle for reasoning beyond the unique situated actions that assessment performances constitute. They tell us what is important in assessment situations, what to look for in students' work, and the terms in which we can say something about examinees that holds meaning beyond the immediate performance. In short, every element of an assessment argument and of the activities that bring it to being can be traced back to this conceptualization. In an assessment argument cast in sociocognitive terms, we look for evidence that a student has been able to marshal resources to act in productive ways in the situations defined by targeted practices or LCS patterns. This could be anything from automatized arithmetic operations to intellective competence. We should design the situations, the performances, and the evaluations to suit both the conceptualization of the capabilities and the purposes of the feedback loops they are meant to support.

Alternative explanations are intimately related to contemporary views of test validation (Bachman, 2005; Cronbach, 1988; Kane, 1992, 2006; Messick, 1989, 1994). Observations are viewed, claims are made. What else could have led to the observations other than the favored claim? The following section will relate alternative explanations to contextualization (i.e., additional information) and standardization.

So far we have sketched out how the argument metaphor applies to assessment design and interpretation. The argument shown in Figure 3 ends at the top with the claim that is based on assessment performances. In practice, the claim is only a way-station to real-world action: A

claim is, in its own turn, data in an argument for assessment use, namely whatever instructional, selection, licensure, or any other purpose the assessment is meant to support. **Error! Reference ource not found.** shows how the assessment design argument can be extended to assessment use (Mislevy, 2006, 2008a). The extended framework is useful for studying the reasoning from assessment practices to intended uses. This activity is test validation in the "assessment as measurement" metaphor, and studying effectiveness and appropriateness of information from the "assessment as feedback loop" metaphor. (See Bachman (2005) and Bachman and Palmer (2010) for in-depth discussions of assessment use arguments.)

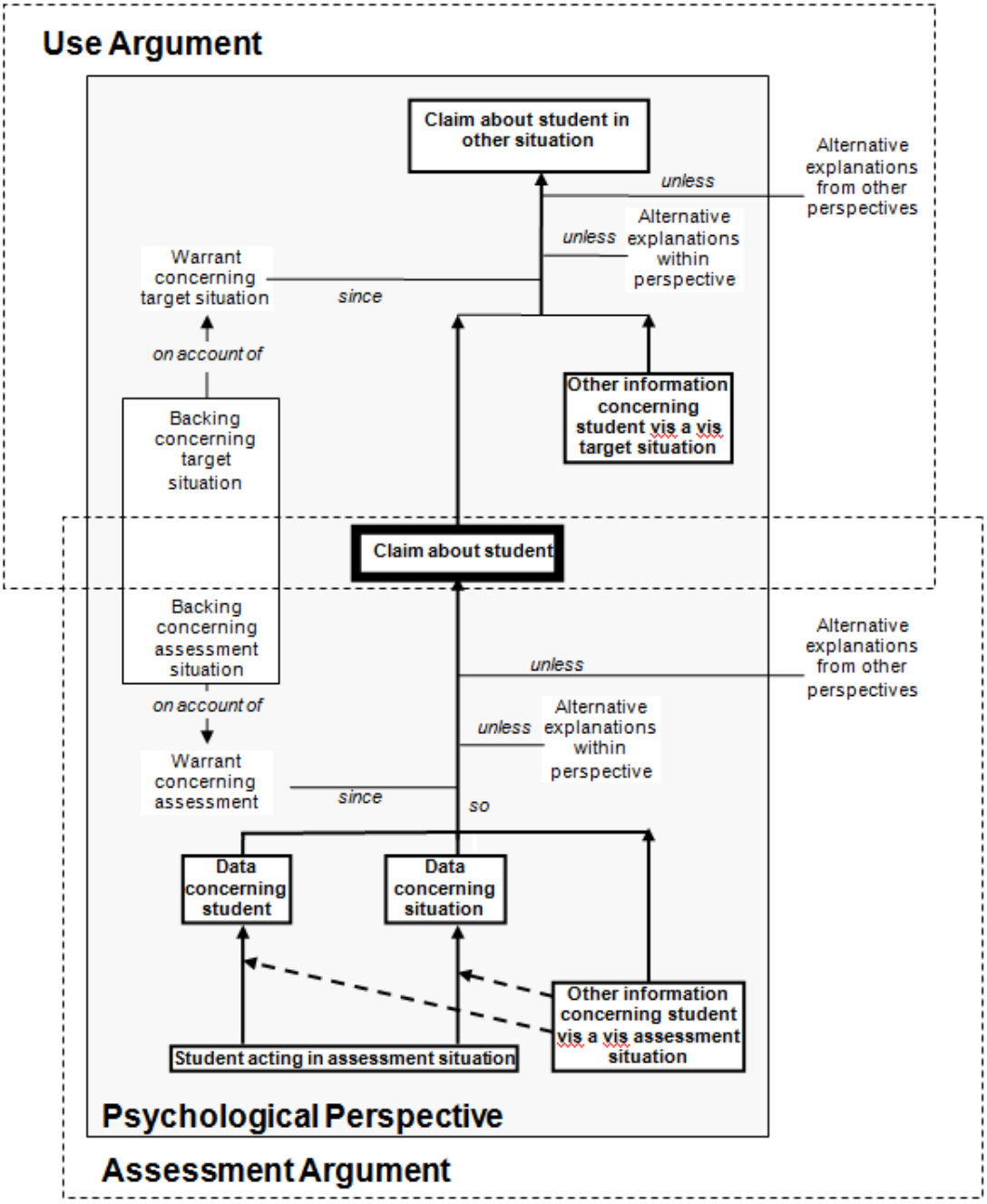


Figure 5. Elaborated structure for assessment arguments. Lower rectangle shows assessment design argument; upper rectangle shows assessment use argument. They share psychological perspective, backing, and claims about student based on assessment.

5.2 Implications

The argument metaphor emphasizes that an argument is made from the knowledge standpoint of some person or group. When information about context is available, the user can incorporate it into her interpretation of performance and have more precise inferences. Such information undercuts alternative explanations for good or poor performance that would arise had this information not been available. These are alternative explanations concerning what the student has studied or other relevant aspects of her learning history, such that the user who does not have this information must take them into account and which downgrade the evidentiary value of the observations.

Standardization is also a design strategy that can mitigate alternative explanations, although with counterbalancing costs. For example, raters might tend to give higher scores to longer essays, or ones that students spent more time on, or ones that incorporated feedback from English teachers. When we must assess from afar, we might therefore decide to standardize on these aspects of a writing assessment. Standardizing on these facets of the task makes comparisons more fair in these respects, since they no longer differ for examinees. On the other hand, we forgo the opportunity to obtain evidence about students' capabilities to write at different lengths and their capacities to write and revise over extended periods of time. These are important issues that we would hope are addressed in the contexts of their writing instruction and assessment locally. An alternative strategy for the feedback involving distant raters would be to permit differences but obtain information about them, which would then need to be incorporated somehow into the evaluation—not easy in this example, but maybe possible in other cases, such as when Olympic competitors choose which dive to do but scores are adjusted by the difficulty of the dive.

Gee's Halo and Algebra comparison provides a second example. Suppose Ms. Strahan learns that Carlo has completed the Carnegie Learning algebra course, and that Jane, who she doesn't know, has completed some course that is called algebra. If she is familiar with the Carnegie Learning course, Ms. Strahan can infer much more about Carlo's capabilities than about Jane's. Jane might have completed the Carnegie course too, but maybe a better one, or maybe a worse one, or maybe one that emphasizes different skills and ways of using them. The existence of these possibilities attenuates her inference. She can obtain some additional data by administering a standardized assessment to both students. The new information is not timed or

tuned to make decisions in tight feedback loops as they study, and it will be a less complete picture of what each student is able to do in various situations than more tailored assessments could give—but it does provide added evidentiary value to Ms. Strahan. Common, standardized assessment is a tool that enables an external user to determine what the information will target (e.g., state standards), within the severe constraints that this kind of assessment must meet. Inferences about more abstract qualities such as intellective competence would almost certainly require greater use of context. Additional information will be required to craft or to recognize situations in which insight or efficacy is required from a given student, and how these qualities are to be evaluated in performances. The same performance in the same situation can reflect these qualities to different degrees for different students. A compromise can be standardizing some aspects of a performative assessment but not other tasks, as is done in AP Studio Art. Rather than standardizing the specifics of the performative situation, such assessments specify higher-level features of activities that must be satisfied and higher-level qualities of performance that must be exhibited. As in AP Studio Art, it may even be part of the assessment to require the examinee to make the case that she has meet the standards of performance. This feature itself provides an opportunity for students to learn the qualities of good work (Wolf, Bixby, Glenn, & Gardner, 1991).

Recall that AP Studio Art is designed to support feedback loops at two levels: tighter, contextualized feedback for learning within each of hundreds of art classes, and larger feedback loops for decisions about awarding college credit for demonstrated accomplishments in the AP course. We can elaborate the assessment argument to show how different actors at different levels in a system can be responsible for different phases of reasoning. Mislevy (2008a) shows a number of configurations to do this. Figure 6 is the configuration that corresponds to AP Studio Art portfolio assessments.

A number of tools and representations have begun to appear that help designers craft assessments from first principles using the assessment as argument metaphor. The evidence-centered design (ECD) framework proposed in Mislevy, Steinberg, and Almond (2003) is the particular strand of research from which the Toulmin argument diagrams shown here were developed. This work is proving helpful for developing new forms of assessment such as simulation-based and game-based assessments (e.g., Behrens et al., 2004; Shute, 2011). One representational form that can prove useful to developing assessments of abstractly defined

capabilities that need to be contextualized (like intellective character) is the assessment design pattern (Liu & Haertel, 2011). Design patterns draw on research and experience to suggest a design space for building contextualized assessment tasks around higher-level capabilities, organized around attributes that correspond to the elements of Figure 3. For example, Mislevy, Riconscente, and Rutstein (2009) provide design patterns to help construct model-based reasoning assessments in science.

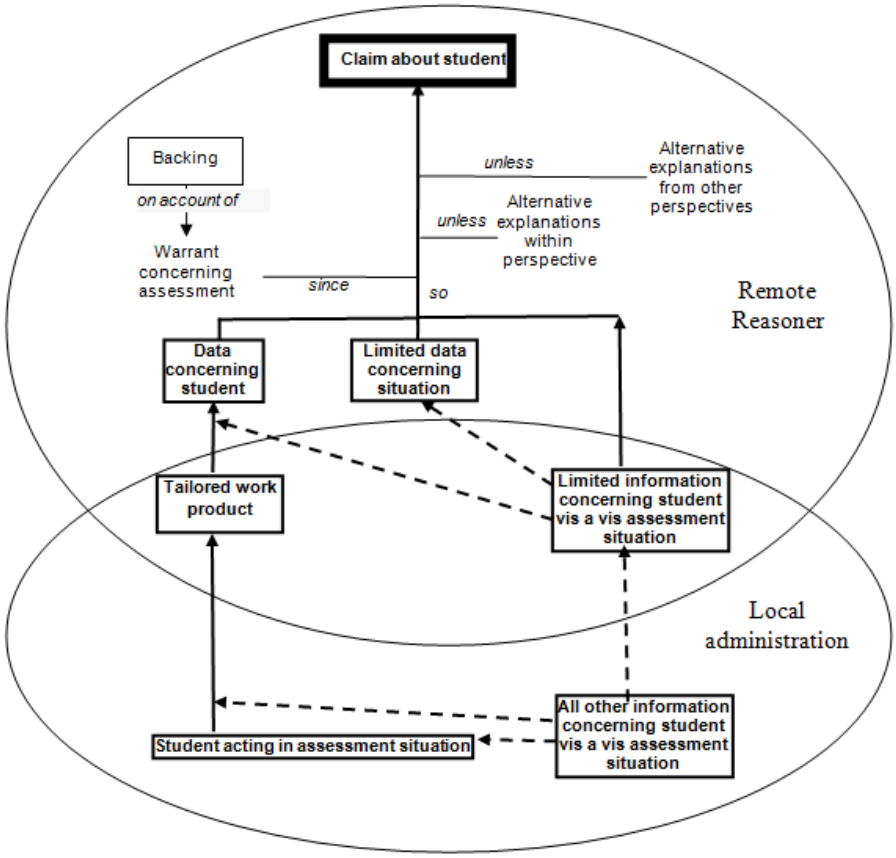


Figure 6. Toulmin diagram for an assessment argument, showing a remote reasoner with locally tailored work products and assessment context and some limited contextual information for remote interpretation.

6. Assessments as Measurement

Measurement is the most familiar of the four metaphors we suggest as a basic toolkit. In fact, it is so familiar that it is not always even recognized as a metaphor; "educational measurement" is often considered to be synonymous with "educational assessment." This section highlights four key points for thinking about assessment with the measurement metaphor. The first point is that it actually is a metaphor. Measurement as a philosophical concept builds up from the direct physical experience of comparing objects in terms of their length or height, for example. Formalizing these simple foundations leads to more abstract concepts of measuring procedures for physical properties, then derived properties such as acceleration, axioms for measurement relationships, and the even more abstract relationships in conjoint measurement in social sciences (Michell, 1999). In the late 1800's psychophysicists Fechner and Weber applied the paradigm to sensory capabilities such as detecting light and characterizing loudness. Psychometrics combines measurement models with statistical models for dealing with noisy evidence. From its origins in the early 1900's, the variables in educational and psychological measurement models, such as intelligence and reading comprehension ability, were considered real characteristics of people with quantitative properties with the same ontological status as force and mass (Michell, 1999). Although this view is still advanced by some psychometricians (e.g., Borsboom, Mellenbergh, & van Heerden, 2004), an alternative view of measurement models is metaphorical—specifically, that fitting and using measurement models in assessment is an instance of model-based reasoning (Mislevy, 1997, 2008a, 2009).

Psychometrics is the application of statistical theory to inference about individuals' capabilities (Mislevy, 1994). It is possible to apply these models as symbol-system frameworks for reasoning about patterns in peoples' performance in situations. These patterns arise not because people have inherent values of traits, but because of the regularities in LCS patterns and practices, and regularities in the ways people tend to learn them and develop resources as they participate in them. This view encourages considerable caution on the part of the modeler: The model is not seen as a representation of a grounded truth, but a provisional frame for reasoning about patterns which could differ under different circumstances (e.g., students who learn a different strategy, as in Tatsuoka, Linn, Tatsuoka, & Yamamoto, 1988). This is the juncture between the practice metaphor and the measurement metaphor. We are thus cautious about reasoning through the model in general, and alert to response patterns that may indicate that a

particular student does not accord well with the typical patterns in a data set. In terms of the argument metaphor, a measurement model is part of the warrant through which we reason. It requires backing in terms of global fit and rationales, and using it opens us to alternative explanations for the performances of individuals with atypical patterns of performance.

The second point is that there is a lot more to measurement models than test scores and reliability coefficients. Even classical test theory and its cousin generalizability theory provide a large variety of tools for characterizing patterns in data, quantifying evidence and flagging anomalies, constructing and improving tests, and designing assessment systems. The AP Studio Art program uses measurement models, for example, not to "measure art proficiency" but to monitor patterns of ratings across a widely distributed system of performances and judgments, across raters, portfolios, and students, to pop out atypical instances that signal errors or unique performances (Myford & Mislevy, 1995). Beyond classical theory are latent variable models that provide flexible methods for modeling multiple aspects of performance in complex situations. The goals are to synthesize evidence across multiple and possibly varied observational situations, and characterize both the strength of evidence and magnitudes of uncertainty for tempering inferences. Measurement models provide a quantitative framework to augment the qualitative arguments discussed in the preceding section.

The third point is that the use of measurement models is not bound to inference cast in trait and behavioral psychology, even though they originated under those paradigms. One can reconceive the patterns being modeled as emerging from phenomena at the cognitive layer shown in Figure 1, and the goal is examining them through the lens of patterns and practices at the social layer. Using measurement models does not presuppose a simplistic realist view of the variables for students that appear in the models. The machinery that evolved under these paradigms does, however, provided statistical models for managing evidence and uncertainty in a probability framework. The framework can be gainfully applied, with appropriate re-interpretations and extensions, to assessment cast in information-processing and sociocognitive perspectives on learning (Mislevy, 2006, 2008b).

The fourth point is that the use of measurement models is not bound to data from familiar forms of assessment. Interest is burgeoning in more complex and interactive forms of assessment made possible by advances in technology—assessment in the digital ocean, as DiCerbo and Behrens (2012) put it, as opposed to the digital desert. Rather than a handful of data points,

performances can be continuous, provide massive amounts of data about both processes and products, can be embedded in interactive simulation environments, may have multiple students interacting with one another, and can have learning occurring during the course of observation. Learning scientists and psychometricians have been collaborating to develop measurement models to characterize such data. As examples, see Williamson, Mislevy, and Bejar (2006) on automated scoring of complex performances, Romero, Ventura, Pechenizkiy, and Baker (2011) on data mining in rich digital assessment performances, Koenig, Lee, Iseli, and Wainess (2010) on measurement models for game and simulation environments, and Shute (2011) on unobtrusive psychometric modeling in games.

The bottom line is that the underlying principles of reasoning from uncertain evidence that lie underneath measurement models are not bound to either the epistemology or the technology under which the paradigm evolved. The same principles can be brought to bear, in some cases reinterpreted and in other cases extended, for radically new forms of assessment.

7. Four Additional Metaphors

The preceding sections reviewed what I consider to be a minimally sufficient set of metaphors for understanding assessment. This section touches on four more that bring a sharper focus to various aspects of assessment. The first two specialize aspects of the practice, feedback, and evidentiary reasoning metaphors. The latter two pull hard in opposite directions on sociocultural aspects of the practice metaphor.

7.1 Tests as Contests

Paul Holland (1994) contrasted "tests as contests" with "tests as measurement." When tests are used as contests, examinees are competing against one another for scarce resources, such as jobs, college admission, or certifications. Thinking of tests as contests calls attention to incentives for examinees to maximize their chances. They may prepare in ways that favor higher scores at the expense of better learning, or they may cheat. This highlights the responsibilities of test creators to create and maintain fair contests. Within the limits of resources, they should produce tests for which better learning is in fact likely to produce higher scores. They should

produce reliable outcomes obtained through transparent scoring of tests administered under clear rules, and they should strive to minimize cheating (Dorans, 2011).

This metaphor contextualizes certain high-stakes tests, and underscores the rights and responsibilities of examinees and examiners in the system. It can thus be considered as highlighting a special case within the "assessment as practice" metaphor. The particular issues Holland examines result from the competitive nature of tests in certain feedback loops, and the consequences it entails for actors in the system. Holland wanted to show how measurement machinery can be brought to bear to help achieve social values such as reliability, validity, comparability, and fairness.

7.2 Assessment Design as Engineering

Familiar methodologies for constructing assessments, such as test specifications, curricular analysis, and rules of thumb for good item-writing practice, serve well for constructing familiar kinds of tests for familiar purposes. Tasks are independent, they each provide a score, and item scores add up to test scores which can then be analyzed using classical test theory. These tools are not up to the job of designing and analyzing assessments that are more complex, in senses such as the following: interactive performances, drawing on multiple aspects of knowledge and skill in different configurations, tasks are inter-related, learning occurs during the course of the experience, the task must be defined in part by the examinees themselves, or examinees can collaborate. Further, issues of cost can exacerbate design challenges when advances technologies are employed.

A current line of work is developing concepts, tool, and processes to improve the quality and efficiency of task design. Embretson illuminated a path toward integrating test theory, task design, and psychology with her 1985 edited volume *Test design: Developments in psychology and psychometrics*. Her (1998) cognitive assessment design system, Luecht's (2003) integrated approach to test design, development, and delivery, and Mislevy, Steinberg, and Almond's evidence-centered design framework (2003) are examples of what Luecht calls "assessment engineering." This metaphor applies the principles of design under constraint (Simon, 2001) to machinery from the evidentiary argument and measurement metaphors, for creating and implementing assessments. This approach is proving especially useful for developing efficient

and valid technology-based assessment, such as through simulations and games (e.g., Behrens et al., 2004; Shute, 2011).

7.3 Examination as the Exercise of Power

Michel Foucault (1977) wrote of assessment as the exercise of power.[9] Modern societies exercise power by making individuals visible and "normalizing them." Examinations are a mechanism for doing this. McNamara and Roever (2006) applied this sociocultural (as opposed to sociocognitive) perspective to the charged topic of language testing, showing how tests function at times as "weapons within situations of inter-group competition and conflict" (p. 196). Similarly, economists view professional licensure examinations through the lens of "barriers to entry." This metaphor is valuable not merely for detecting overt abuses of authority, but for sensitizing us to more subtle but pervasive shaping effects that various assessment practices have on individuals and groups. This metaphor connects with the measurement metaphor through the all-important concept of validity, specifically with respect to the controversial role of consequences in test validation (Messick, 1989).

This metaphor returns us to the policy strategy mentioned in Section 3 of using assessment as a lever for change. The idea starts with an existing educative system of some kind, whether it is a national educational system, a licensure regime, or a classroom and grading policy. The system has an assessment component, which shapes peoples' behavior within the system as the system exists, both as they understand it and in ways they are not aware of; it signals what is important to learn, and influences how learners learn, teachers teach, and evaluators evaluate. Change the assessment, the theory goes, and changes ripple through the system (Resnick & Resnick, 1992). This is an application of the "assessment as feedback loop" metaphor, applied at the level of the system itself. Sometimes change does occur, and sometimes in the ways that are intended. Success depends in part on the scale and complexity of the system. Large, dispersed, or complex systems are harder to change, as are ones where the intended change disadvantages stakeholders who have levers of their own for stasis.

[9] As he thought of almost everything.

7.4 Assessment as Inquiry

Constructivist educators chafe at the 'exercise of power' metaphor. Its positive counterpart is the "assessment as inquiry" metaphor, in which assessment is a principled but open way of interacting with students to produce and evaluate evidence about the capabilities they are developing, interacting with others and their communities, in ways and in directions that are not tightly determined by authorities beforehand (Delandshire, 2002; Serafini, 2000):

> We are moving here from an educational practice of assessment where we have defined a priori what we are looking for, to an educational practice where we are participating in activities in which we formulate representations to better understand and transform the world around us. If our purpose is to understand and support learning and knowing and to make inferences about these phenomena, then it seems that the idea of inquiry—open, critical, and dialogic—rather than that of assessment —as currently understood) would be more helpful, as it would encourage consideration of the epistemological and theoretical assumptions from which we work. (Delandshire, 2002, p. 1475)

Of particular interest would be to operationalize "assessment as inquiry" in terms of evidentiary arguments, given that the arguments would be more individualistic and less constrained than the pre-formed arguments that familiar assessments are built around. My view is that they can be, given that the evidentiary framework has evolved as much from the individualistic and open settings of jurisprudence as it has from science and statistics (Schum, 1994), and has been applied as such to the evaluation of portfolios in studio art (Myford & Mislevy, 1995). The evidentiary framework provides concepts to reveal analogues in structure and reasoning between such different forms as standardized tests, simulations, and portfolio assessments, as well as to highlight their differences and understand the reasons for them.

8. Conclusion

Discussions of assessment reform are complicated by the fact that there are many kinds of assessments, used in different ways for different purposes. Moreover, many disciplines are involved in assessments of various kinds for various uses. A change that improves assessment

practice in one way for one purpose might impair it for another. Thinking and talking about assessment, much less setting assessment policy, can scarcely afford to take place at the level of surface features.

This paper provides four powerful metaphors to organize thinking about assessment in all its guises. It is clearly beyond the present scope to fully describe, illustrate, and connect them, in a variety of contexts and practices. It is unlikely that any single person could provide such an analysis, as each metaphor connects to deep and subtle bodies of knowledge that would take a lifetime of study to master individually. But awareness of the key ideas from each and their interconnections provides some common language to begin investigations of particular assessments, of assessment's roles in systems, and in the future of assessment. These metaphors make us aware of conceptual frameworks we can take advantage of, and hook us into the experiences, the tools, and the wisdom of a many disciplines. The metaphors do not resolve questions about assessment, but they do help us ask them sensibly.

References

Bachman, L.F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2,* 1-34.

Bachman, L.F., & Palmer, A. (2010). *Language assessment in practice.* Oxford, UK: Oxford University Press.

Behrens, J. T, Mislevy, R. J., Bauer, M., Williamson, D. W., & Levy, R. (2004).Introduction to Evidence Centered Design and lessons learned from its application in a global e-learning program. *International Journal of Testing, 4,* 295-301.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Boulet, J.R., Smee, S.M., Dillon, G.F., & Gimpel, J.R. (2009). The use of standardized patient assessments for certification and licensure decisions. *Simulations in Healthcare, 4,* 35-42.

Braun, H.I., & Mislevy, R.J. (2005). Intuitive test theory. *Phi Delta Kappan, 86,* 488-497.
Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test Validity* (pp. 3-17), Hillsdale, NJ: Erlbaum.

Delandshire, G. (2002). Assessment as inquiry. *Teachers College Record, 104,* 1461-1484.

DiCerbo, K. E. & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz (Ed.) *Computers and their impact on state assessment: Recent history and predictions for the future.* Charlotte, North Carolina: Information Age Publishing.

Dillon, G.F., & Clauser, B.E. (2009) Computer-delivered patient simulations in the United States Medical Licensure Examination (USMLE). *Simulation in Healthcare, 4*, 30–34.

Dorans, N. J. (2011). *The contestant perspective on taking tests: Emanations from the statue within.* Invited 2010 Career Award address at the annual meeting of the National Council on Measurement in Education (NCME), April 7-11, 2011, New Orleans, LA.

Embretson, S.E. (Ed.) (1985). *Test design: Developments in psychology and psychometrics*. Orlando: Academic Press.

Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Fillmore, C. J. (1976). Frame semantics and the nature of language. In S. R. Harnad, H. D. Steklis, & J. Lancaster (Eds.), *Origins and Evolution of Language and Speech* (pp. 20-32). Annals of the NY Academy of Sciences, Vol. 280.

Foucault, M. (1977). *Discipline and punish: The birth of the prison.* New York: Vintage Books.
Frezzo, D.C., Behrens, J.T., & Mislevy, R.J. (2009). Design patterns for learning and assessment: facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *The Journal of Science Education and Technology.* Springer Open Access http://www.springerlink.com/content/566p6g4307405346/

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24,* 355-392.

Gordon, E.W. (2008). Intellective competence. *The Oracle, 85*(17), 21-22, 28.

Greeno, J.G. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53*(1), 5-26.

Holland, P. W. (1994). Measurements or contests? Comments on Zwick, Bond and Allen/Donoghue. In *Proceedings of the Social Statistics Section of the American Statistical Association,* 27–29. Alexandria, VA: American Statistical Association.

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus & Giroux.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527–535.

Kane, M. (2006). Validation. In R. J. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 18-64). Westport, CT: Praeger.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* New York: Cambridge University Press.

Koenig, A. D., Lee, J. J., Iseli, I., & Wainess, R. (2010). A conceptual framework for assessing performance in games and simulations. *CRESST Report 771.* Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press

Lemke, J. (2000) Across the scales of time: Artifacts, activities, and meanings in ecosocial systems. *Mind, Culture, and Activity 7,* 273-290.
http://ecd.sri.com/downloads/ECD_TR11_DP_Supporting_Task_Authoring.pdfLiu, M., & Haertel, G. (2011). Design patterns: A tool to support assessment task authoring. *Large-Scale Assessment Technical Report 11.* Menlo Park, CA: SRI International. Available online at http://ecd.sri.com/downloads/ECD_TR11_DP_Supporting_Task_Authoring.pdf

Luecht, R. M. (2003). Multistage complexity in language proficiency assessment: A framework for aligning theoretical perspectives, test development, and psychometrics. *Foreign Language Annals, 36*, 527–535.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* Oxford: Blackwell.

Melnick, D. (1996). The experience of the National Board of Medical Examiners. In E.L. Mancall, P.G. Vashook, & J.L. Dockery (Eds.), *Computer-based examinations for board certification* (pp. 111-120). Evanston, IL: American Board of Medical Specialties.

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

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

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* New York: Cambridge University Press.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R.J. (1997). Postmodern test theory. In A. Lesgold, M. J. Feuer, & A. M. Black (Eds.), *Transition in work and learning: Implications for assessment*, pp. 180-199. Berkeley, CA: McCutchan.

Mislevy, R.J. (2003). Substance and structure in assessment arguments. *Law, Probability, and Risk, 2*, 237-258.

Mislevy, R.J. (2006). Cognitive psychology and educational assessment. In R.L. Brennan (Ed.), *Educational Measurement* (Fourth Edition) (pp. 257-305). Phoenix, AZ: Greenwood.

Mislevy, R.J. (2008a). Issues of structure and issues of scale in assessment from a situative/sociocultural perspective. In P. A. Moss, D. Pullin, E. H. Haertel, J. P. Gee, & L. J. Young (Eds.). *Assessment, equity, and opportunity to learn* (pp. 259-294). New York: Cambridge University Press.

Mislevy, R.J. (2008b). How cognitive science challenges the educational measurement tradition. *Measurement: Interdisciplinary Research and Perspectives, 6,* 124. Available online at http://bearcenter.berkeley.edu/measurement/docs/CommentaryHaig_Mislevy.pdf

Mislevy, R.J. (2009). Validity from the perspective of model-based reasoning. In R.L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 83-108). Charlotte, NC: Information Age Publishing.

Mislevy, R.J., Riconscente, M.M., & Rutstein, D.W. (2009). Design patterns for assessing model-based reasoning *PADI-Large Systems Technical Report 6*. Menlo Park, CA: SRI International. Available online at http://ecd.sri.com/downloads/ECD_TR6_Model-Based_Reasoning.pdf

Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

Myford, C.M., & Mislevy, R.J. (1995). Monitoring and Improving a Portfolio Assessment System. *CSE Technical Report 402*. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Norman, D.A. (1993). *Things that make us smart.* Boston: Addison-Wesley.
Razzaq, Feng, Heffernan, Koedinger, Nuzzo-Jones, Junker, Macasek, Rasmussen, Turner & Walonoski (2007). Blending Assessment and Instructional Assistance. In Nadia Nedjah, Luiza deMacedo Mourelle, Mario Neto Borges and Nival Nunesde Almeida (Eds). *Intelligent Educational Machines within the Intelligent Systems Engineering Book Series*. pp.23-49. (see http://www.isebis.eng.uerj.br/). Springer: Berlin / Heidelberg.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A., Rasmussen, K.P. (2005). The Assistment Project: Blending Assessment and Assisting. In C.K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.) *Proceedings of the 12th International Conference on Artificial Intelligence In Education*, 555-562. Amsterdam: ISO Press.

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Conner (Eds.), *Future assessments: Changing views of aptitude, achievement , and instruction* (pp. 37-75). Boston: Kluwer.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. (Eds.). (2011). *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press.

Saxe, G.B. (1988). Candy selling and math learning. *Educational Researcher, 17*(6), 14-21.

Schum, D.A. (1987). *Evidence and inference for the intelligence analyst.* Lanham, Md.: University Press of America.

Schum, D.A. (1994). *The evidential foundations of probabilistic reasoning.* New York: Wiley.

Serafini, F. (2000). Three paradigms of assessment: Measurement, procedure, and inquiry. The *Reading Teacher, 54,* 384-393.

Shaffer, D.W. (2006). *How computer games help children learn.* New York: Palgrave/Macmillan.

Shute, V.J. (2011). Stealth assessment in computer-based games to support learning. In S Tobias, JD Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers, 2011.

Simon, H.A. (2001). *The sciences of the artificial* (4th ed.). Cambridge, MA: MIT Press.
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23,* 645–665.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement, 25,* 301-319.

Toulmin, S.E. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Wertsch, J. (1998). *Mind as action.* New York: Oxford University Press.

Williamson, D.M., Mislevy, R.J., & Bejar, I.I. (Eds.). (2006). *Automated Scoring of complex performances in computer based testing.* Mahwah, NJ: Erlbaum Associates.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Educational Research, Vol. 17* (pp. 31-74). Washington, DC: American Educational Research Association.

Young, R.F. (2009). *Discursive practice in language learning and teaching.* Malden, MA: Wiley-Blackwell