Issues of Structure and Issues of Scale in Assessment From a Situative/Sociocultural Perspective

CSE Technical Report 668

Robert J. Mislevy
CRESST/University of Maryland

January 2006

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-153

Project 3.6 Study Group Activity on Cognitive Validity
Robert J. Mislevy, Project Director, CRESST/University of Maryland

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

ISSUES OF STRUCTURE AND ISSUES OF SCALE IN ASSESSMENT FROM A SITUATIVE/SOCIOCULTURAL PERSPECTIVE[1]

Robert J. Mislevy

CRESST/University of Maryland

Abstract

A situated/sociocultural (SC) view of assessment "emphasizes questions about the quality of students' participation in activities of inquiry and sense-making, and considers assessment practices as integral components of the general systems of activity in which they occur" (Greeno, Collins, & Resnick, 1997, p. 36). This presentation addresses two issues. The first is understanding the SC view of assessment through the lens of an "evidence centered" design framework that has proven useful for assessment cast in trait, behavioral, and information-processing perspectives. The second is addressing issues that arise when one attempts to design assessments that are at once compatible with SC principles and suitable for large-scale use. Illustrations are drawn from the Advanced Placement Studio Art portfolio art assessment and the HYDRIVE intelligent tutoring system.

1. Introduction

A situative/sociocultural (SC) perspective "views knowledge as distributed among people and their environments, including the objects, artifacts, tools, books, and the communities of which they are a part. Analyses of activity in this perspective focus on processes of interaction of individuals with other people and with physical and technological systems" (Greeno, Collins, & Resnick, 1997, pp. 16-17). Accordingly, a "situative view of assessment emphasizes questions about the quality of students' participation in activities of inquiry and sense-making, and considers assessment practices as integral components of the general systems of activity in which they occur" (Greeno et al., 1996, p. 37). Research on school learning from the SC perspective "incorporates explanatory concepts that have proved useful in fields such as ethnography and sociocultural psychology to study collaborative work…mutual understanding in conversation, and other

[1] This work was also supported by the Spencer Foundation's "Idea of Testing" project. I am grateful to my Idea of Testing colleagues for stimulating discussions on the issues addressed herein, to Lyle Bachman for discussions on test use arguments, and to Drew Gitomer and Lyle Bachman for comments on an earlier draft.

characteristics of interaction that are relevant to the functional success of the participants' activities" (Greeno et al., 1996, p. 37). In such analyses, attention focuses on patterns of interactions that occur in detailed and particular situations, yields "thick" descriptions of the activities, and often produces voluminous data. Studies at this level of detail are essential for understanding the conditions and the interactions through which students learn; that is, "opportunities to learn" that particular circumstances afford particular students, in light of their particular personal and educational histories of experience.

However, no practical assessment at the level of the classroom, let alone a school or a program, can demand scores of hours of videotape per student, all analyzed by a team of graduate students, each producing a multipage ideographic report. Methodologies for micro-level SC analyses and for large-scale assessment must differ, to be sure, but what about explanatory concepts? Is an SC perspective irreconcilable with the very idea of large-scale assessment? Or are there methods and concepts at another level of explanation that can be used, different from but compatible with SC explanations, in the sense that Boyle's law is compatible with the motions of individual molecules of a gas?

An understanding of assessment that is based solely on experience with large-scale standardized testing might suggest the answer is no. One sees decontextualized tasks, dissociation from classroom activities, and statistical models originally conceived to answer questions cast in trait and behaviorist psychology. Yet while these testing practices are familiar and widespread, recent advances in technologies, methodologies, and practical needs have given rise to forms of large-scale assessment practices with two key characteristics: They are compatible with an SC perspective in the "levels of explanation" sense of the previous paragraph, and to accomplish this they draw upon methods and concepts that have arisen out of a psychometric tradition, but have been extended or reconceived as necessary to support SC interpretations.

The following presentation argues this case. It uses the "evidence centered" assessment design (ECD) approach described in Mislevy, Steinberg, and Almond (2002), to illuminate the structure of assessment arguments and assessment design frameworks. The ECD structures have been used for analyzing and designing assessments cast in terms of trait, behavioral, and information-processing psychological perspectives (e.g., Mislevy et al., 2002). It is posited that the same structures hold value for analyzing and designing assessments cast in terms of an SC

perspective as well, with the meanings of the elements of the ECD structures appropriately construed.

To this end, we begin by briefly reviewing Toulmin's (1958) structure of arguments, then specializing it to assessment arguments. The central role of the psychological perspective is emphasized. It grounds the interpretation of every element in the argument—the nature of claims one wishes to make about students' learning; interpretations of the things they say, do, or make; and the observational situations in which they act and interact. While the focus is on assessment arguments from an SC perspective, similarities and contrasts with assessment under trait, behavioral, and information-processing perspectives prove useful. We will see that knowledge about the interrelationship among students, their histories, and assessment contexts plays a larger role in SC assessment, and presents accordingly greater inferential challenges for persons further from the assessment context in detail, time, and distance.

High-level representations of models for the formal assessment structures typically used in large-scale assessment settings are then presented. After recalling their uses and meanings in familiar testing practices, we examine reconceptualizations that, in a compatible assessment system, would support interpretations consistent with a SC perspective on learning. The assessment argument and design structures help bring out the ways that the situativity of knowledge and the contextualization of interpretation are dealt with in large-scale assessment systems.

These ideas have been put into practice in several places, and a number of them are noted here. The two that play the largest role in the discussion are the Advanced Placement Studio Art portfolio art assessment (Mitchell, 1992) and HYDRIVE (Gitomer, Steinberg, & Mislevy, 1995), an intelligent tutoring system to help Air Force trainees learn to troubleshoot the hydraulics system of the F-15 aircraft. AP Studio Art blends situated classroom practice and large-scale, high-stakes assessment: Work 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, share, and critique pieces. HYDRIVE is based on information-processing principles, but functions as a learning tool in ways consistent with sociocultural principles and can be used to support decisions cast in trait and behaviorist terms.

2. Assessment as Argument

Philosopher Stephen Toulmin (1958) proposed a schema for how we use substantive theories and accumulated experience to reason from particular data to particular claims. Figure 1 outlines the structure of a simple argument. The *claim* (C) is a proposition we wish to support with *data* (D). The arrow represents inference, which is justified by a *warrant* (W), a generalization that justifies the inference from the particular data to the particular claim. Theory and experience—both personal and formal, such as empirical studies and prior research findings—provide *backing* (B) for the warrant. In any particular case we reason back through the warrant, so we may need to qualify our conclusions because there may be *alternative explanations* (A) for the data. Alternative explanations will themselves be supported or undercut by *rebuttal data* (R). This section extends Toulmin's structure to assessment arguments.

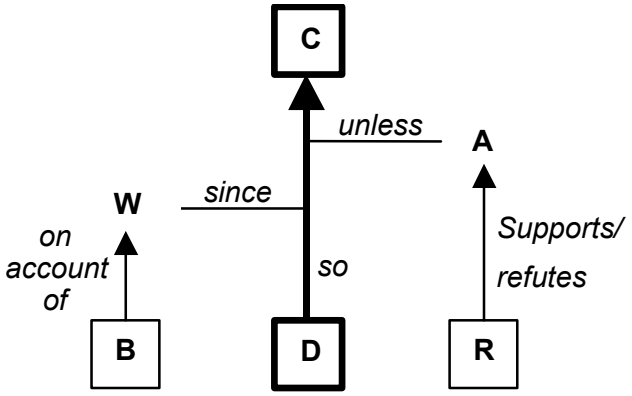


Figure 1. Toulmin's (1958) structure for arguments. Reasoning flows from *data* (D) to *claim* (C) by justification of a *warrant* (W), which in turn is supported by *backing* (B). The inference may need to be qualified by *alternative explanations* (A), which may have *rebuttal evidence* (R) that tends to support or refute support them.

2.1 The Relevance of a Perspective on Knowledge and Learning

The foundation of an educational assessment argument is a conception of the nature of proficiency. A psychological perspective shapes the nature of all the elements in the argument structure and the rationale that orchestrates them as a coherent argument. What kinds of things might one wish to say about persons (claims)? What kinds of things does one need to have a person say or do in what

kinds of situations (data)? How are they related (warrants)? What is observable is a person's action—actually a constellation of actions, indeed interactions, with elements of the environment and sometimes other people, in some social context. But there are countless aspects of persons, of situations, and of persons' actions within situations, to which we might attend, and countless ways we might characterize them. A conception of proficiency shapes what among these we will perceive, and which will constitute data in a given assessment argument.

Discussion is facilitated by using terms from four stereotypical psychological perspectives for thinking about knowledge and learning (adapted from Greeno, Pearson, & Schoenfeld, 1997, and Greeno et al., 1996). They differ as to which of the aspects of human learning, thinking, acting, and interacting they bring to the foreground, and consequently in terms of the nature and instantiation of assessment arguments cast in their light.

- *A behaviorist perspective*. The behaviorist psychological perspective focuses on targeted behavior in a domain of relevant situations. Details of both the behavior and the situation, as construed by the observer, are in the foreground; internal mechanisms and representations are moved to the background, even rejected as unscientific in the strictest versions of the perspective. Knowledge is viewed as the organized accumulation of stimulus-response associations, developed and strengthened through reinforcement from the environment, that serve as components of more broadly defined skills.

- *A trait or differential perspective.* Messick (1989) defines a trait as "a relatively stable characteristic of a person—an attribute, enduring process, or disposition—which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances (p. 15)." People learn many different things and act in many different situations, not just from one person to the next, but from one time and situation to another for the same person. Variables intended to hold meaning across people over time may be proposed to characterize consistencies within individuals, evidenced as systematic differences among individuals. From the trait perspective, test scores hold value to the extent that behaviors observed in the assessment context are manifest in some context of use, despite differences between the contexts' demands for knowledge of particular content, tools, and social situations. Also in the background for the trait perspective are mechanisms that produce behavior and the conditions of learning that precede it.

- *An information-processing perspective.* Epitomized in Newell and Simon's (1972) *Human Problem Solving*, the information-processing perspective examines the procedures by which people acquire, store, and use knowledge to solve problems. The focus is on "what's happening within people's heads"—not just what a person does in a situation as seen from the outside, as under a behavioral perspective, but in terms of the patterns—the meanings—through which a person perceives, construes, and interacts with a situation. Parallels with computation as symbol manipulation play an important role in the information-processing perspective, in the use of rules, production systems, task decompositions, and means-ends analyses.

- A *situative/sociocultural* perspective. Much learning is motivated and shaped by the knowledge, goals, constraints, and physical presence of other people. Social organizations such as families, classrooms, professions, and so on, influence the processes of acquiring, storing, representing, understanding, and creating knowledge. These influences are channeled by particular ways of communicating: genres, conventions, knowledge representations, and so on. "Sociocultural" highlights the activities through which knowledge is created, conditioned, constrained, and brought to bear, in the contexts of the technologies, information resources, representational forms, and social systems that constitute the situations in which people act. "Situative" highlights how people construct tailored and specific meanings to each new situation around patterns from past experiences, in each instance modifying and extending the repertoire of patterns and experiences they can bring to bear in the next situation.

Of course neither learning nor assessment can be partitioned neatly into discrete bins with these labels. The problems, the interfaces, and the feedback in HYDRIVE, for example, are all built around the information-processing notions of defining an active path in a problem space, carrying out test procedures, and applying strategies such as space-splitting and serial elimination. But the ways HYDRIVE is used reflect a sociocultural perspective. This includes problem solving in pairs or small groups to promote communication in terms of the language of troubleshooting, and scaffolding for trainees that decreases as they become more proficient—"cognitive apprenticeship" in the manner of Collins, Brown, and Newman (1989). With feedback turned off, the same simulator can be used to estimate the proportion of problems in the domain a trainee can solve to support a decision about whether he is ready for the flightline or should continue training. Here we see an assessment purpose and assessment procedure cast in behavioral terms, in concert nevertheless with the information-processing and sociocultural grounding of the training system in which it is embedded.

2.2 The Structure of Assessment Arguments

Figure 2 is an extension of Toulmin's (1958) structure to assessment arguments. Although still quite simplified, it incorporates features that help one understand similarities and differences among assessment arguments cast in different psychological perspectives. It is not difficult to relate this structure to formal and familiar assessments because their visible parts and processes are set up explicitly before the assessment occasion, and they map fairly directly to elements of the argument. This is the topic of Section 3. But the same structure could be used to analyze a conversation between a student and a teacher as they work through, say, making sense of a poem—in this case with the arguments implicit, constructed on the fly, reconstructed iteratively moment by moment as new actions are observed and new meanings are made by all involved.

Figure 2 actually distinguishes two main arguments, the assessment argument *per se* in the lower dashed rectangle (Mislevy, 2003) and an assessment use argument in the upper rectangle (Bachman, 2005). Our attention will focus on the assessment argument, but assessment cannot be understood apart from purpose and use. Recognizing the flow from assessment data to assessment use, it is useful nevertheless to distinguish the mediating structure of claims about students in order to understand the role of psychological perspective.

The assessment claims are shown in the center of the figure as output of the assessment argument and data for the use argument. They are the terms in which we organize, summarize, and understand observations made in the assessment setting, for subsequent reasoning in the use setting. They connect our thinking about what is observed in assessment settings with our thinking about assessment purposes such as guiding, evaluating, and affording students' learning; evaluating, improving, and monitoring instructional systems; and selecting, placing, and assigning individuals to opportunities. The *meaning* of the mediating claim is thus integral to both *perception* of student's actions in the assessment situations and subsequent *action* in the use situation, all consistent under the guiding perspective. Practically all of the elements of both arguments are circumscribed in the box labeled "psychological perspective," to emphasize how each is construed through that perspective. Alternative explanations are an exception. Some alternative explanations that we need to ameliorate or take into account rise within the psychological perspective that guides the assessment design project. But others can

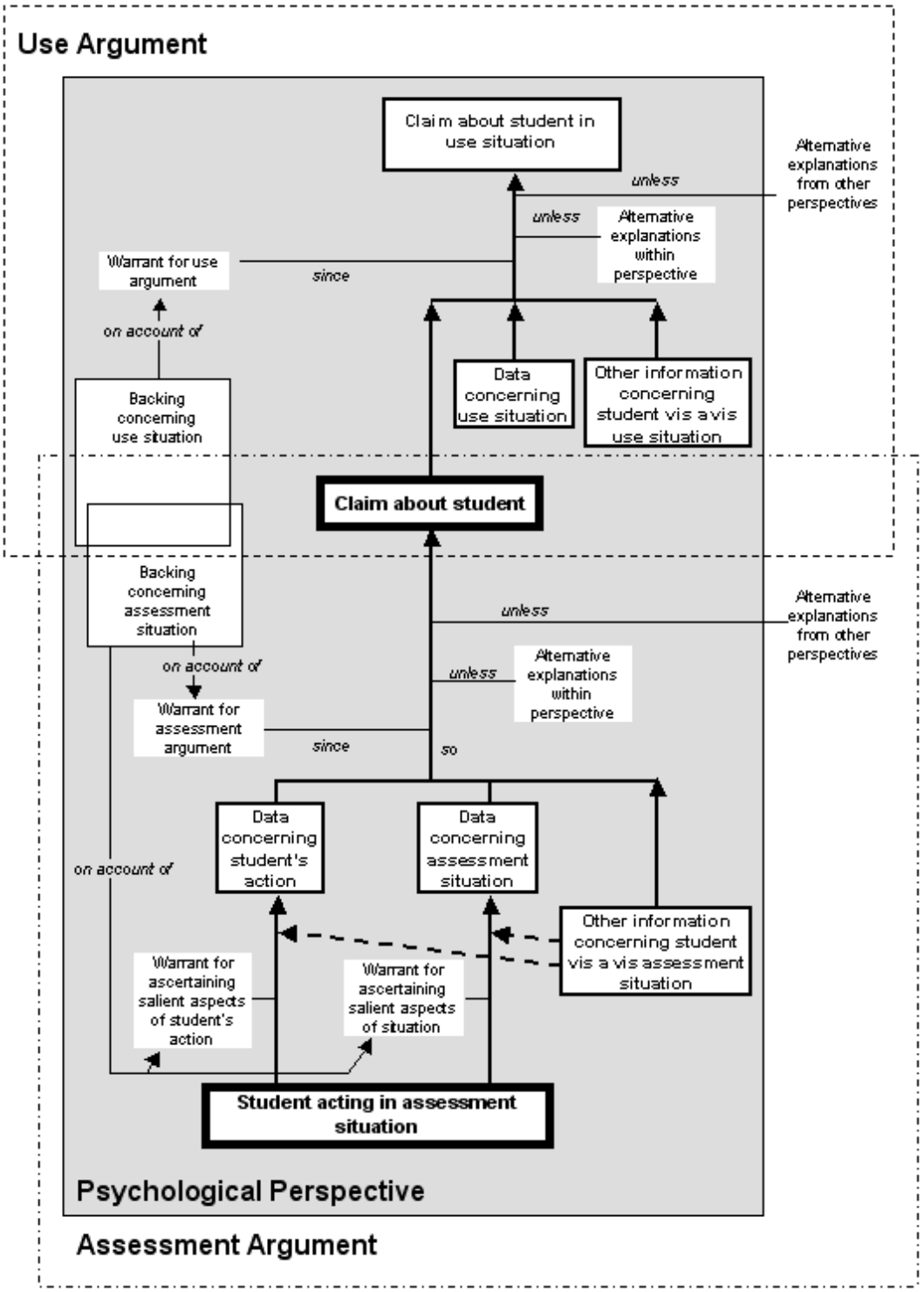


Figure 2. Elaborated structure for assessment arguments. Lower rectangle shows assessment argument proper; upper rectangle shows assessment use argument. They share psychological perspective, backing, and claim about student based on assessment.

arise from other perspectives, cast in terms of entities, relationships, or explanations that lie outside the narrative space the guiding perspective.

2.3 What Are Data?

With regard to the role of psychological perspective in assessment, a particularly interesting part of the argument structure in the lower box concerns the data that ground the claim about the student. The nature of data, and the actions and situations from which data arise, are driven by the nature of the claims, which are cast in terms of psychological perspectives. Highlighted at the bottom left of Figure 2 is a student's actions in a situation, the unit of analysis in assessment: The student says, does, or makes something, possibly extending over time, possibly interacting with others. Interpretations of the actions rather than the actions themselves constitute data in an assessment argument. Note that warrants are required for these interpretations, cast in terms of the psychological perspective and the substantive grounding of the argument. These paths of argumentation from situated actions to data will be expanded in Section 3.3, as they can exhibit multiple steps and be carried out by different actors with different responsibilities, points of view, or bodies of information. At this first pass, though, we see that an assessment argument generally encompasses three kinds of data:

- aspects of the situation in which the person is acting,
- aspects of the person's actions in the situation, and
- other information about the person's history or relationship to the observational situation. This information may be further required to interpret the action in the situation, to interpret the situation as it applies to the particular person, or through which to interpret the aforementioned kinds of data as they pertain to the claim.

Aspects of the situation and the action in the situation. The first two of these are characterizations of aspects of a person acting within a situation that might hold value beyond the single, unique event. More fully, they are understandings of aspects of actions in particular assessment situations that could help one understand other events that have happened or to anticipate events that might happen—with regard to the same person acting in a different situation, past, future, or hypothetical; or, with regard to the situation, what might happen if other persons, similar or different in defined ways, were to interact with situations that are similar

in defined ways. Such inferences are needed for educative planning, for a student in the first instance, for designing instruction in the second.

The most common understanding of "assessment data" is characterizations of students' actions in assessment settings—famously, right or wrong answers, but more broadly, characterizations of qualities in open-ended performances, use of models or strategies, contributions to and interactions in group projects, and so on. Aspects of situations are left in the background, having been thought about by test developers but not seen as part of the data proper. Yet inference from assessment settings to use settings depends critically on a theory (perhaps implicit) of situations. One must make the case that features of the assessment settings reflect features of the targeted use situations that elicit the relevant knowledge, skill, or propensities, however conceived. Principled design of assessment tasks—which features of use settings are relevant and critical, which add realism at the expense of introducing demands for irrelevant knowledge, for example—thus requires an understanding of features of situations, in light of a conception of the proficiencies that are of interest in those situations (Messick, 1994).

These issues have received particular attention in language testing, where the targeted language use often involves complex uses of language in complex situations (Bachman & Palmer, 1996; Douglas, 2000). Just how one can simplify and standardize situations to meet practical constraints, yet still obtain evidence about inherently social and interactive capabilities, is ever a challenge. To this end, background research for proposed revisions of the Test of English as a Foreign Language (e.g., Enright et al., 2000) included insights into the pragmatic and social situations in which people use language (sociolinguistics) as well as results on the nature and acquisition of language (psycholinguistics). Figure 2 explicitly indicates that this backing grounds both the assessment argument and the use argument. The same backing grounds the warrants for interpreting aspects of students' actions and the situations. It may be the case that the theory upon which these evaluations are based requires more than simply observing the action in the situation. Additional knowledge may also be required to condition these judgments as they are used to create assessment settings and to evaluate actions within them.

The role of other information. "Other information" data are essential to assessment arguments, even though they are often tacit, embedded in forms and practices. What we know about a particular student acting in a particular situation can influence how we interpret the aspects of the interaction that will constitute data

about the person and the situation. This possibility is indicated by the dashed lines from the "Other Information" box to the lines rising from student's action to "data concerning the student's actions" and "data concerning the situation."

The kinds of additional information that may be required, and the implications for inference that result if it is missing, vary across assessments framed under different psychological perspectives. As a simple illustration, the American Council on the Teaching of Foreign Languages' (ACTFL) reading guidelines (ACTFL, 1989) contrast Intermediate readers' competence with texts "about which the reader has personal interest and/or knowledge" (ACTFL, 1999a ¶ 2) with Advanced readers' comprehension of "texts which treat unfamiliar topics and situations" (ACTFL, 1999b ¶ 3)—a distinction fundamental to their underlying conception of developing language proficiency. If we wish to assess students' proficiency in a foreign language, we must decide how we want to think of proficiency. Suppose, on one hand, the target of inference is cast in behavioral terms, as overall proficiency with respect to a domain of tasks. We can predefine successful behavior on each task the same way for all students regardless of their familiarity, administer a sample of tasks to a student, and thereby obtain direct evidence about expected behavior in the domain. Suppose, on the other hand, the target of inference is level of proficiency through the lens of the ACTFL guidelines. If we know that the context of a given situation is familiar to one student but unfamiliar to a second, the same observed behavior from the two students holds radically different evidential import about their ACTFL levels. Additional information thus conditions the evidentiary value of students' performances. Which of these two conceptualizations of language proficiency is the correct one? This question makes no sense without an assessment purpose in mind. For determining comparative levels of language proficiency and familiarity with a specified knowledge base, then successful performance on random samples from the corpus is appropriate. For determining individual students' proficiencies or the purpose of planning instruction, using texts known to be familiar or unfamiliar to each, and characterizing their proficiency from the ACTFL perspective, is more useful.

Arguments cast in the behavioral perspective move to the background the role of additional information in characterizing both the student's action and the situation. Ideally, any observer would be able to follow the respective evaluation procedures and come up with the same interpreted data, both with regard to characterizing features of the stimulus situation and features of the action.

Bormuth's (1970) linguistic transformation rules for generating a universe of comprehension tasks for a reading passage is an example, with the advertised advantage that any researchers would be led to identical universes of test items based on a given text.

In trait-based arguments, background information comes to the fore for investigating alternative explanations of performance; first, in looking for interactions between performance on tasks and background variables, in the form of test and item bias, and second, in circumscribing the range of background characteristics across which inferences can be made without conditioning interpretations of performance on their values. Procedures that have evolved to examine these questions include differential item functioning analyses (DIF; Holland & Wainer, 1993) and generalizability analysis (Cronbach et al., 1972).

In information-processing arguments, students' prior experience or familiarity with goals, procedures, and representational forms is essential for designing complex performance tasks and then interpreting actions in the resulting situations (Mislevy et al., 2002). Simulation-based task performances, for example, require interpretations across multiple, continuous sequences of actions and interactions, to ground claims about use of strategies, familiarity with affordances, and so on. In HYDRIVE, it is not the particular troubleshooting actions that a student carries out that constitute data, but rather the troubleshooting strategy that the action best accords with in light of the actions the student has taken thus far, and the evolving information they have provided and the changes they have caused to the situation up to that point.

Arguments cast in SC terms generally require the greatest use of additional information in both interpreting students' actions and characterizing the features of assessment situations. Some relevant aspects of situations, such as contexts and materials, can be characterized across students, but other aspects of situations that are necessary to understand a student's actions are aspects as the student perceives them. Similarly, some aspects of students' actions, such as the meter and word choices, can be characterized from just work products, but others, such as whether a style or a phrase extends a structure from a student's family experiences, cannot be recognized without knowing that connection.

Section B of AP Studio Art is the student's "concentration," up to 20 slides, a film, or a videotape illustrating a student-selected theme. An excerpt from Gasser's

(1955) classic text gives a feel for how the Concentration taps into a fundamental aspect of what it means to "be an artist." Gasser discusses the experience of running into a difficulty in drawing or painting a particular subject, and suggests isolating a particular problem and exploring it with a variety of angles:

> This is a procedure that insures progress, and it is one that many professional artists follow. They will work a long time on a single theme—anything from a still life containing a textural problem to nocturnes. It can be subject matter of a religious nature, a scene in a foreign country, or the lighting effect on a particular surface. Whatever the subject, the professional artist makes exhaustive studies of it. When he feels that he has interpreted the subject to the extent of his capabilities he may have a one-man exhibition whose theme is the solution of the problem. It is surprising how few people who view the paintings realize this; most regard it simply as subject matter that has appealed to the artist. This can be partly true, but only the artist knows to what extent he has met the challenge of solving his particular problem. (p. 85)

The work in a concentration is produced over the course of the school year, as students and teachers in each of hundreds of participating schools create, discuss, share, and critique pieces. These interactions are situated with respect to individual students' interests, experiences, and capabilities, and with respect to materials, pieces of work, episodes of creation and discussion. Both the informal assessments represented in ongoing feedback and discussion and the more "official" grades for the work or the course draw on the teachers' in-depth knowledge of local circumstances. Yet these discussions and grades are also shaped by the common requirements by which all portfolios are rated centrally at the end of the year. The generally stated standards are the foundation of the Section B warrants. With every student's unique concentration they must be interpreted anew—by the student and the local teacher interacting in the class, and later by the central raters. The determination of a student's topic, the approach he or she takes, the details of individual pieces, and the evaluation of the work are a matter of negotiation between the teacher and the student throughout the year. This experience is at once necessary for assessment and central to the learning experience that AP Studio Art is meant to provide. How these local assessment/learning interactions are aligned with the common, more limited end-of-the-year evaluations is discussed in Section 3.3.

3. Scaling Up

The argument structure of Section 2 is quite flexible with respect to not only psychological perspectives, but also as to whether it is constructed before or after observations, whether an argument is crafted for each new case or the same framework is used for multiple episodes or students, and how much judgment and how much additional information may be required for intermediate inferences.

Practical work is not so accommodating. Each assessment has purposes to serve and constraints to meet. Just who needs what information, for what use, at what scale, with what costs, and with what implications for learning at the system level? We may distinguish between small-scale assessments used in context to guide learning, which exploit local additional information and support local uses, from assessments in which certain key users are distant from the learning context in terms of time, space, and information. These properties characterize large-scale uses, which have the additional property of needing to make assessment arguments for many students.

How can assessment arguments be scaled up and made portable? What tradeoffs to the qualities of evidence and the validity of inferences result? Do the tradeoffs differentially affect arguments from different psychological perspectives? Four courses of action for designing assessments at large scales and conveying information outside the immediate situation are these:

- using the same argument structure for many students,
- making the machinery—that is, the processes and artifacts by which the assessment argument is effected— formal and explicit,
- structuring the use of information in interpreting assessment situations and students' actions, and, in particular, constraining the use of additional information, and
- using probability-based reasoning to synthesize bodies of evidence and characterize the strength of information they provide for claims.

3.1 Using the Same Argument Structure for Many Students

The assessment argument structure can be applied to classroom quizzes and standardized achievement tests, to coached practice systems and computerized tutoring programs, and to the informal conversations students have with teachers.

In the last of these examples, decisions about kinds of observations, tentative hypotheses, and reasoning from one to the next, are unconstrained and assembled on the fly. In the rest, a framework has been predetermined for the kinds of data that will be gathered, the kinds of claims that will be made, and the rationales that support the inference.

If we foresee that similar data can be gathered for similar purposes on many occasions, we can achieve efficiencies by developing standard procedures both for gathering the data and reasoning from it (Schum, 1994, p. 137). A narrative space is predefined: A general story line, the kinds of claims that can be proposed, the range of data that will support them. The tradeoff is this. On the one hand, a well-designed protocol for gathering data addresses important issues in its interpretation, such as thinking through the kinds and amounts of evidence that are required to support claims, and to head off certain likely or pernicious alternative explanations. The warrant and the backing for many individual arguments can be communicated to the remote user. On the other hand, only those stories that can be framed in the predetermined narrative space can be told.

The term "standardization" associated with testing is best understood in terms of argument structures that are to some degree determined in advance. Standardization concerns the structure of the argument and selected aspects concerning settings, standards, rubrics, representations, instructions, or contexts—and possibly, but not necessarily, the form of the data. We mean to avoid the colloquial identification of standardization with multiple-choice items, independent work, and time limits. There are hundreds of aspects of any assessment that could be standardized or not, to varying degrees, in myriad configurations. They can concern different parts of the assessment argument. Standardization is a strategy for heading off certain alternative explanations for good or poor performance, such as varying amounts of time or support, that could affect students' performance for reasons unrelated to our purposes, and thereby strengthen claims.

Concerning the situations in which students will act, the idea is to foresee what features of the prospective action-within-situation need to be satisfied by the person in the situation in order to satisfy the requirements of the warrant through which inference will be made. That is, at least some of the conditions of the situation are arranged so that *the data concerning the situation* needed in the assessment argument will be applicable. The nature of the features depends in part on the psychological perspective in which the warrant is framed. One can predetermine objective

features of the situation as seen from the assessor's point of view (e.g., circumstances, directives, materials, and affordances provided in the assessment situation), or more generally stated characteristics of the situation that may be determined by the assessor with additional knowledge of the student, chosen by the student under given constraints, or negotiated by the student and the assessor in ways that satisfy generally stated features of the assessment setting. For behaviorist arguments, objective features are all that count. Objective features may usefully be specified in trait, information-processing, and situative/sociocultural arguments as well, but generally stated characteristics are increasingly important, to be determined by specific instantiations that satisfy the generally stated characteristics as they apply to particular individuals and their circumstances. Recall the example of language assessment texts that the assessor knows to be either familiar or unfamiliar to a particular student. A student's topic for her AP Studio Art concentration is an example of a negotiated determination of specifics. Two examples of concentrations are as follows (Myford & Mislevy, 1996):

> My concentration project grew out of a desire to explore angularity in a medium (clay "wheel-work") which doesn't easily permit a graceful, lyrical expression of that term. I was initially intrigued by random geometric shapes depicted on rounded surfaces—often repeated on appendages of the main work—sometimes incised or emphasized by a glazing technique. Recently, I have begun to investigate those same geometric planes literally piercing one another as I have initiated an exploration of metal and wood. Reflective qualities and light(ing) have frequently been a concern as well. (p. 7)

> The subject of my concentration is minimalist oriental landscapes particularly reminiscent of Chinese and Japanese landscapes. My fascination with landscapes and intense color use inspired me to emulate ancient oriental styles along with minimalist simplification of forms and clutter. I utilized their techniques of depicting the serenity of nature through simple yet bold brush stokes and colors. My materials comprised of watercolors and airbrush. My series began with uncomplicated scenery and gradually building on to bolder use of form and color. (p. 7)

Concerning actions of a student within the assessment situation, preconstruction again looks ahead to what kinds of features of actions-within-situations are needed in the argument, and guides or constrains students' actions so that what they say, do, or make can exhibit the relevant qualities. That is, at least some of the conditions of the assessment are arranged so that *the data concerning the student's actions* needed in the assessment argument will be applicable. Of all the activity in the assessment setting, certain expectations are made clear to the student

as to the form of the performance that is expected, the qualities it should exhibit, and the (possibly overlapping) qualities in terms of which it will be evaluated. Specific work products may be defined as the agreed-upon trace of action that will constitute the body of evidence to be evaluated—anything from vectors of multiple-choice responses, to keystroke-level traces of actions in HYDRIVE, to videotapes of teaching classroom lessons in teacher certification examinations. Again the nature of the features depends in part on the psychological perspective in which the warrant is framed. And again one can predetermine objective features of the performance as seen from the assessor's point of view (e.g., selection of alternatives, successful repair of a fault in the hydraulics system, completion of the required number and form of pieces in an AP Studio Art concentration), or more generally stated characteristics of the performance that may be determined by the assessor with additional knowledge of the student, chosen by the student under given constraints, or negotiated by the student and the assessor in ways that satisfy generally-stated features of the targeted performance. For behaviorist arguments, objective features are all that count. Objective features may usefully be specified in trait, information-processing, and situative/sociocultural arguments as well, but generally stated characteristics are increasingly important, to be evaluated in specific performances in accordance with the generally stated characteristics as they apply to particular individuals and their circumstances.

Concerning procedures for evaluating students' performances, again procedures are predetermined in specifics or in general terms to be later specified, as may be required to suit the warrant that justifies inference in the assessment argument. Procedures for evaluating students' work products are typical in large-scale assessments. The specified procedures could be automated or require human judgment. The more complex performances are, however, the more important it becomes that students understand the qualities and criteria the evaluation procedures embody. Again as one moves away from behavioral arguments, this is a critical link in not only the assessment argument but the learning. In HYDRIVE, understanding that space-splitting in a problem space is a positive feature in evaluation is a facet of understanding what space-splitting is and recognizing when to do it. In assessments such as AP Studio Art and teacher certification examinations, coming to understand the evaluation procedures is integral to learning goals: "[Q]uestions of what is of value, rather than simple correctness … an episode in which students and teachers might learn, through reflection and debate,

about the standards of good work and the rules of evidence" (Wolf, Bixby, Glenn, & Gardner, 1991, p. 51).

Few large-scale assessments are less standardized in the traditional sense than the Advanced Placement Studio Art portfolio assessment. Students have an almost unfettered choice of media, themes, and styles. But the AP program provides considerable information about the qualities students need to display in their work, what they need to assemble as work products, and how raters will evaluate them. This allows for a common argument, and heads off alternative explanations concerning unclear evaluation standards.

Predetermining all or some links in an assessment argument, then preconstructing assessment elements and prearranging procedures to effect those links offers efficiencies, but it admits the possibility of cases that do not accord with the common argument. The assessor thus acquires two responsibilities: To establish the credentials of the evidence in the common argument, and to detect individuals for whom the common argument does not hold. Inevitably, the theories, the generalizations, and the empirical grounding for the common argument will not hold for some students. These instances call for additional data or different arguments, often on a case-by-case basis.

Predefining the narrative space does not specify the psychological perspective underlying that space, but the implications of this constraint are felt more sorely under an information-processing perspective than under a behavioral or trait perspective, and even more under a situative/sociocultural perspective. One loses, it would seem, tailored arguments, thick descriptions, and 'emic' (as opposed to 'etic') claims. And at the level of the distant user of large-scale assessment results, this is generally true. The final AP Studio Art portfolio scores that colleges use to award credit or waive prerequisites are simply numbers on a 0–5 scale. As discussed in Section 3.3, however, the rating process first entails multiple emic (if brief) evaluations of each portfolio; raters' then map from their constructed understandings of a body of work to numeric summaries of the performance in terms of a common framework of evaluation (Myford & Mislevy, 1996). To ensure coherence with the situative/sociocultural perspective on learning, it is necessary that these private evaluations are cast in the same public framework of meaning that underlies the dispersed classroom interactions. The probability-based models used in AP Studio Art evaluations and discussed in Section 3.4 contribute to this end.

3.2 Making the Machinery Formal and Explicit

External forms of knowledge representation support distributed cognition, or people working together on tasks that are large, complex, extend over time and space, and use specialized information from multiple sources. These adjectives apply to large-scale assessment. Good knowledge representations embody key entities and relationships in a domain, and help people plan and conduct their work in concert with the fundamental principles of the domain. The student-model, evidence-model, and task models in the evidence-centered approach to assessment design (ECD) proposed by Mislevy, Steinberg, and Almond (2002) are meant to serve this purpose (Figure 3 gives a high-level view of the central models, omitting internal structure and details). These models provide schemas for processes, protocols, and artifacts in educational assessments, for planning assessments that embody an assessment argument as described in Section 2.

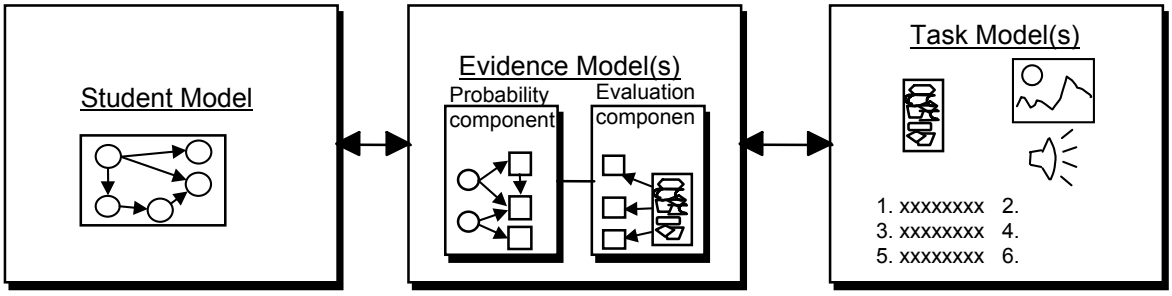


Figure 3. High-level view of central models of evidence-centered assessment design.

In brief, the *student model* specifies the variables in terms of which we wish to characterize students. It is most closely related to the claims in a Toulmin argument structure. *Task models* are schemas for ways to get data that provide evidence about students. Task models specify circumstances of observation and students' work products, both of which are involved as forms of data in the Toulmin structure. *Evidence models* consist of two components which are links in the chain of reasoning from students' performances to their knowledge and skill: The *scoring component* contains procedures for extracting the salient features of student's performances in individual task situations—i.e., ascertaining the values of observable variables—and the *probability component* contains machinery for updating beliefs about student-model variables in light of this information. The scoring component concerns the reasoning from students' actions to the salient aspects thereof. The probability

model concerns synthesizing these data, possibly across multiple tasks, in terms of belief about students, as caricatured in terms of "student model variables." (As discussed in Section 3.4, "student model variables" are better thought of as vehicles for summarizing a reasoner's observations than as properties of students *per se*.) In informal assessments, this component corresponds to inferences about a student in some terms that rise above the particulars of the performance. In formal assessments, the variables in the student model are the observable variables that are related through probability models (also discussed in Section 3.4).

A more fully detailed representation that can be used for operational work expresses these structures in terms of an object model and a corresponding equivalent XML specification (Riconscente, Mislevy, & Hamel, in press). Filling in the schema as appropriate to an assessment argument explicates the elements needed to make the assessment operational. The goal of standards and protocols is to have structures that maximizes sharing while minimizing constraints on the content and meaning of what is shared, much as routers can move packets of information from one computer to another over the internet without regard to the content of the message as the user sees it, be it text, numbers, music, images, or political tracts with diametrically opposed positions. In structures for assessment elements, just how those elements are fleshed out and what meanings they will acquire in use depend on the assessment argument, which may be cast in any of the psychological perspectives discussed previously. For assessments cast under different perspectives, the models and variables can have similar formal structures but very different situated meanings. They are alike in some ways, such as the roles they play in argument structures and connections they have with other elements of the assessment, but they differ as to the meanings derived from the nature of the data and the claims they are meant to support—much in the way that words acquire situated meaning in contexts ("The coffee spilled, get the mop" versus "The coffee spilled, get a broom" versus "The coffee spilled, stack it again;" Gee, 2003).

Insights from HYDRIVE and AP Studio Art suggest two ways that explicit structures can facilitate designing larger scale assessments that are consonant with situative/sociocultural considerations. First, the "mechanical" elements can be shared more efficiently. Second, the articulation between assessment arguments and the elements of operational assessment reveals how activities and contexts impart meaning to the elements, and those meanings are (well, should be) driven by purposes and perspectives rather than by processes and forms.

A student model variable under a behaviorist assessment might stand for the probability that a student will produce the targeted response to a randomly selected stimulus condition in a behavioral domain. The data that constitute evidence are observers' evaluations of actions, made as objectively as possible, in situations structured as objectively as possible to meet the requirements of the stimulus situation description in the behaviorist warrant. An example is successful repairs of hydraulics system faults in HYDRIVE, to determine whether a trainee is ready for the flight line. Note that a behaviorist assessment argument serves here a useful purpose and is concordant with a learning environment cast in information-processing and sociocultural terms.

But the claim space and supporting-data space are not sufficient for the purpose of helping a trainee who is not doing well to improve. Assessment cast in an information-processing perspective is needed (Steinberg & Gitomer, 1996), with finer grain student model variables keyed to practice modules that address facets of declarative, procedural, and strategic knowledge. Sociocultural and situative considerations remain in the background, as the assessment is embedded within the particular technological and social training environment. The meaning of the student model variables is situated in this context by construction. Are students' values on these variables, reflecting as they do actions within the context, useful for trait-style inferences for other purposes and contexts, such as predicting performance on the flight line or proficiency with different aircraft? The information-processing research upon which they are based provides some backing to suggest they may be, in terms of similarities in the reasoning structures that are required across contexts; similarities in affordances and social situations of use offer backing from a sociocultural perspective. Empirical validity studies for the trait-based predictions would be required, though, to provide more fully satisfactory backing for trait-based inferences of this sort.

AP Studio Art portfolio final scores are obtained through the use of psychometric models that were developed for behavioral and trait-based assessment. Yet their situated meanings emerge from the system of learning, producing work, and rating performances. The challenge students and teachers face during the course of the year, and the challenge the central raters face at the end of the year, is to create situated meanings for common standards for quite different behaviors in different contexts—yet in a way that is generally agreed upon as valid and fair. As noted above, one student's concentration focused on "angularity in

ceramics," while another's dealt with an "application of techniques from traditional oriental landscapes to contemporary themes." It would be easier to compare students' performances if everyone were required to work with angularity in ceramics or oriental landscape, or a prespecified sample of topics. But these ways of determining the assessment context provide no opportunity to obtain evidence about conceptualizing and realizing one's own artistic challenges. How well the ceramics student might have fared with oriental landscapes is not directly relevant to the claim of interest. What does matter, and what AP Studio Art must examine the fidelity of, is inference about the more abstractly defined qualities that should be evinced in any student's chosen concentration. The emergent meanings of final numeric ratings in AP Studio Art, then, are neither as estimates of proficiency in a domain of behaviors (a behaviorist perspective) nor as measures of qualities inherent in students (a trait perspective). They are, rather, summary evaluations of particular achievements in contexts crafted to help students learn both techniques and ways of thinking in art (a situated/sociocultural perspective).

3.3 Structuring the Use of Information in Interpreting Situations and Actions

The preceding section discussed how pre-structuring spaces of claims and data is one way to scale up, at the cost of flexibility in interpretation. This section considers approaches to pre-structuring data interpretation that allow some degree of contextualization that is particularly important in arguments cast in SC terms.

Figure 4 accommodates the situation of an observer of a student acting in an assessment situation, and in real time and interactively noticing salient aspects of action and situations as they unfold, constructing claims, re-examining action and situation anew, noticing new aspects, revising claims, and so on. This is how teachers informally assess their students as they interact in small groups, for example, to see how each student is developing ways to communicate mathematical ideas as they solve problems in groups. Figure 4 blows up the assessment argument portion of Figure 2, and includes an oval that represents the purview of this "local reasoner." Everything the teacher knows about students, their histories, and their relationships to the situation and to each other is available for fashioning claims and interpreting actions and situations as they unfold. It can be the case that claims and data interpretations are developed jointly with students, as in the daily interactions in AP Studio Art classrooms. In this fully connected environment, one can construct and instantiate assessment arguments from any psychological perspective—in

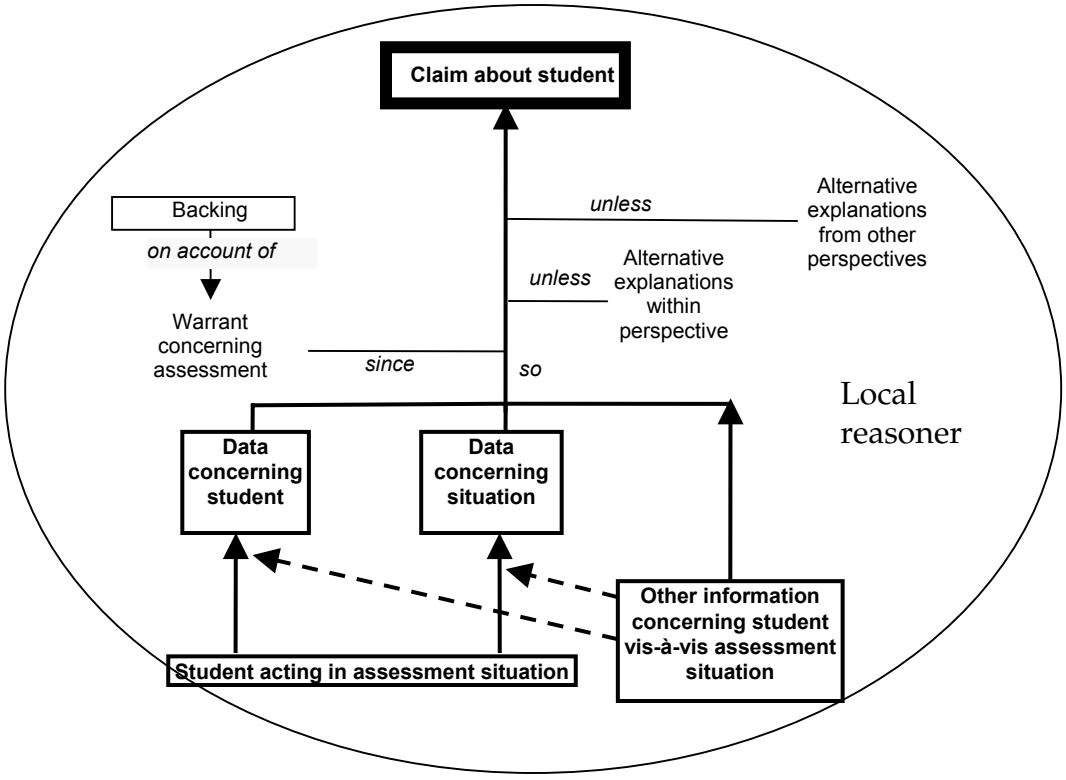


Figure 4. Toulmin diagram for an assessment argument, showing the purview of a local reasoner in a fully connected environment with all argument elements in play.

particular, those from an SC perspective that demand individualized interpretations of actions in situations.

Of course just because someone is doing assessment in a fully connected environment does not guarantee the inferences are good. In most domains, novices differ from experts by not always knowing what to look for, how to interpret what they see, and what to do next (Salthouse, 1991). Teaching is no exception. Student assessment is one of the standards for accomplished practice that the National Board for Professional Teaching Standards (NBPTS) addresses in its portfolio assessment for NBPTS certification. The preparation material for preparing a portfolio of one's practice in Career and Technical Education, for example, asks the candidate if their portfolio will be able to "present evidence of how you use assessment of student work to support learning goals, to facilitate students' growth as career and technical education students, and to inform and shape your teaching practice?"[2]

[2] Downloaded from the National Board for Professional Teaching Standards website on April 2, 2005: http://www.nbpts.org/candidates/guide/whichcert/08EarlyYoungAdult2004.html

Inexperienced teachers can have difficulties because they don't have a good understanding of how students learn, don't have sufficient familiarity with the learning domain itself, and don't know how to interpret students' actions or shape situations that will provide clues about students' understanding.

In contrast to an unconstrained assessment in a fully connected environment, a teacher can give a test with well-defined tasks, with features predetermined to evoke evidence about some targeted capabilities, to be completed individually and evaluated on the basis of features of prespecified work products alone. Figure 5 illustrates this situation. Everything is still under the purview of the local reasoner (i.e., the teacher), but the contextual information does not play a role in determining the data about the student's performance; the evaluations follow predetermined procedures, anything from key matching to human judgment into a common framework. Each of these links could be more fully detailed as Toulmin diagrams in their own right, with warrants, outcomes as claims, and alternative explanations. Generally contextual information does play a role in determining the data about the situation, however; although the tasks are predefined, the choice of these tasks at this time is motivated by a knowledge of where the students are in their course of learning and options for further learning that can be informed by the evidence the tasks will evoke. The teacher's inferences are also conditioned by this information.

Consider the reasoner who is distant from the assessment episode, or who must deal with hundreds or thousands of assessment episodes. It is not possible to carry out tailored argument construction and observation in a fully connected environment (Figure 4). Even with prestructuring, this reasoner must limit the information he or she works with, or reason with data that summarize more contextualized evaluations from local reasoners. The quality of the local evaluations becomes an issue to the remote reasoner: How can one gauge quality without knowing what information was used or the reasoning process that led to the summary?

To outward appearances, the most common way of scaling assessment up looks very much like the procedures described above for the contextualized use of prestructured situations and interpretations in the classroom. The data interpretation phases of large-scale "drop in from the sky" tests are shown in Figure 6. The targeted space of claims is predetermined, as are features of tasks that are

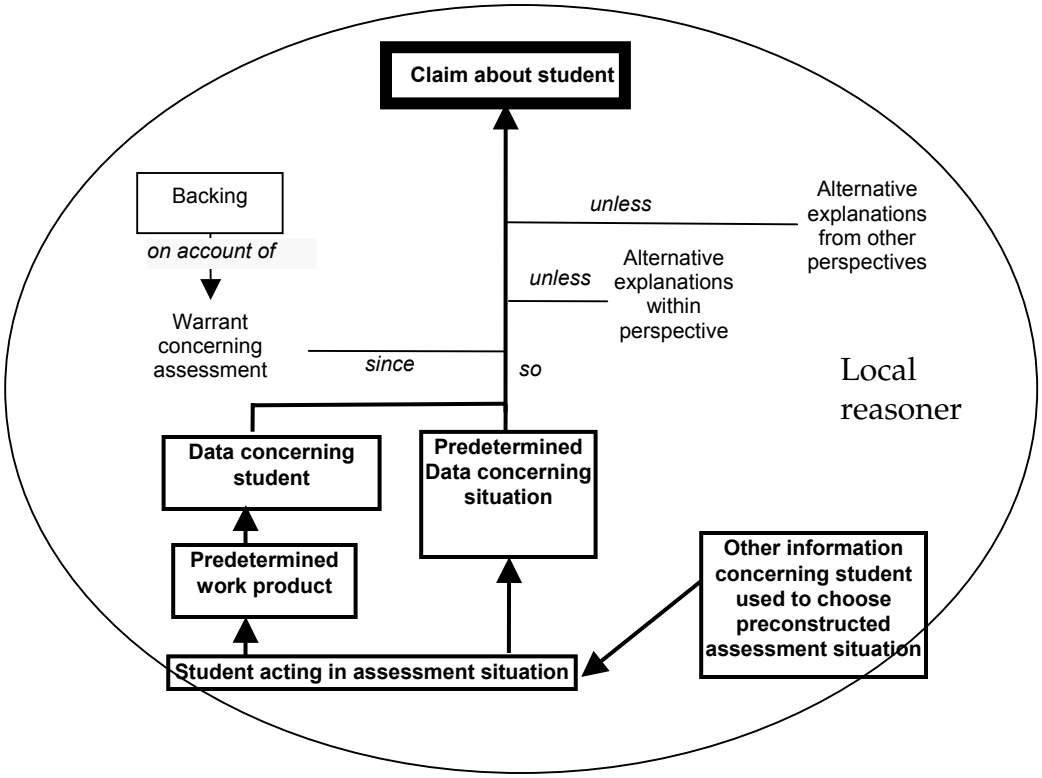


Figure 5. Toulmin diagram for an assessment argument, showing the purview of a local reasoner with predetermined work products and assessment situation, and no links with contextual information in interpretation.

meant to elicit evidence to support the claims, specifications of student work products, and evaluation procedures. The features of tasks are known to the remote reasoner (perhaps they were crafted to evince, for example, national science standards). What is missing is the contextualization of the tasks with respect to students' instructional and personal histories. Even if the same evaluations of work products are derived from the same performances, their meanings for the remote reasoner differ from those of the local reasoner. The space of claims that can supported, and the space of interpretations of the situated actions available to support claims, are both more constrained. Less information is used, but less information is needed to accompany the data for a distant user to know the conditions and procedures that led directly to the data in hand. For arguments cast in behavioral terms, the constrained claim and data space may be fully sufficient. As one moves to trait, information-processing, then sociocultural arguments, the same data provide less satisfactory evidence to ground the claims of interest; too many alternative explanations accord with the observations.

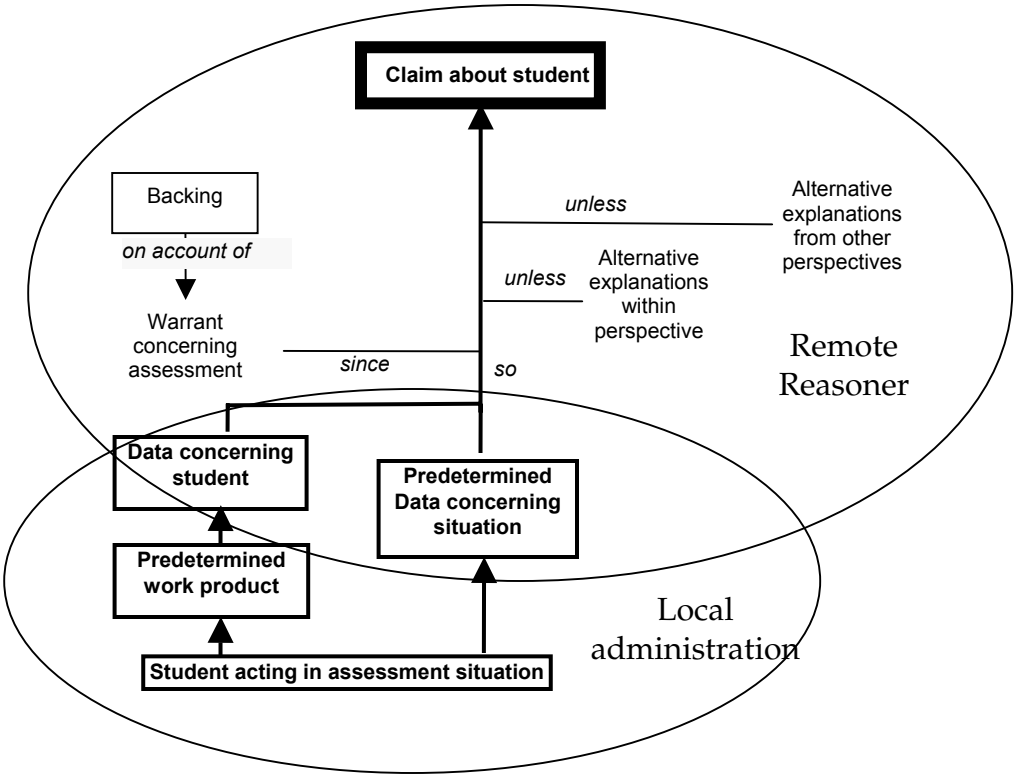


Figure 6. Toulmin diagram for an assessment argument for a remote reasoner, showing locally gathered data with work products and assessment situation preconstructed and no links with contextual information in interpretation.

Figures 7 and 8 represent two configurations midway between the fully connected local environment for reasoning (Figure 4) and the configuration for "drop in from the sky" assessment (Figure 6).

Figure 7 is the approach taken by AP Studio Art. A work product (e.g., the pieces in the Concentration section of a portfolio) is provided to the distant reasoner, the myriad details of its genesis and execution stripped away. But the evaluation of the work and the aspects of the situation in which the work was conceived and carried out are summarized in written explanations that accompany the portfolio. The student submits not only the pieces but paragraphs describing the concentration, relating it to the standards, and discussing the student's goals, intentions, influences, and other factors that help explain the series of works. This material helps the raters figure out just what it was the student had in mind when producing the series of works in her concentration. This is effectively an opportunity for the student to negotiate how the necessarily general principles

expressed in the rubrics should be applied to her particular work. In Section 3.2 we stressed how it was important that the student situate the meaning of the standards in her own work, to serve the goals of learning cast in sociocultural terms. Here we stress how it is important that the rationale for this situated meaning be communicated to the distant reasoner (the central raters), to insure coherence between local understanding with system-wide understanding.

Figure 8 is an alternative approach for utilizing contextual information locally in a large-scale assessment system. Here the data evaluations are done locally, possibly using additional contextual information, by the local reasoner. The distant reasoner obtains summary evaluations, but not the additional information about the situation and the relationship between the student and the situation, which may be integral to the local evaluation. How can the distant observer gauge the value of the local evaluations? Social networks, shared examples, and workshops, as employed in AP Studio Art, all help. More formal strategies include audits (Resnick, 1997), shared benchmark performances and interpretations, and semi-contextualized evaluations across localities, by which local applications of standards can be adjusted to comport better with system-wide evaluations of comparable work. The "social moderation" schemes for adjusting state assessments in Australia reflect the last of these strategies (Linn, 1993).

3.4 Using Probability-Based Reasoning

Toulmin (1958) offers no recipe for characterizing the degree of belief we should assign to claims in a data-based argument, or combining evidence across multiple, possibly overlapping or conflicting, pieces of data. Probability-based reasoning supports coherent reasoning from data to claims, specifically through Bayes' theorem. We may construct a probability model that approximates the key features of the situation in terms of variables and their interrelationships. Although probability-based models can be constructed for unique situations (Kadane & Schum, 1996, do so for the 395 pieces of evidence in the Sacco-Venzetti trial), it is more common in assessment to preconstruct probability models.

There is an important difference between the variables in a probability model and the corresponding entities, claims, and data, in a Toulmin diagram. A claim in a Toulmin diagram is a particular proposition that one seeks to support; a datum is a particular proposition about an aspect of an observation. A variable addresses not

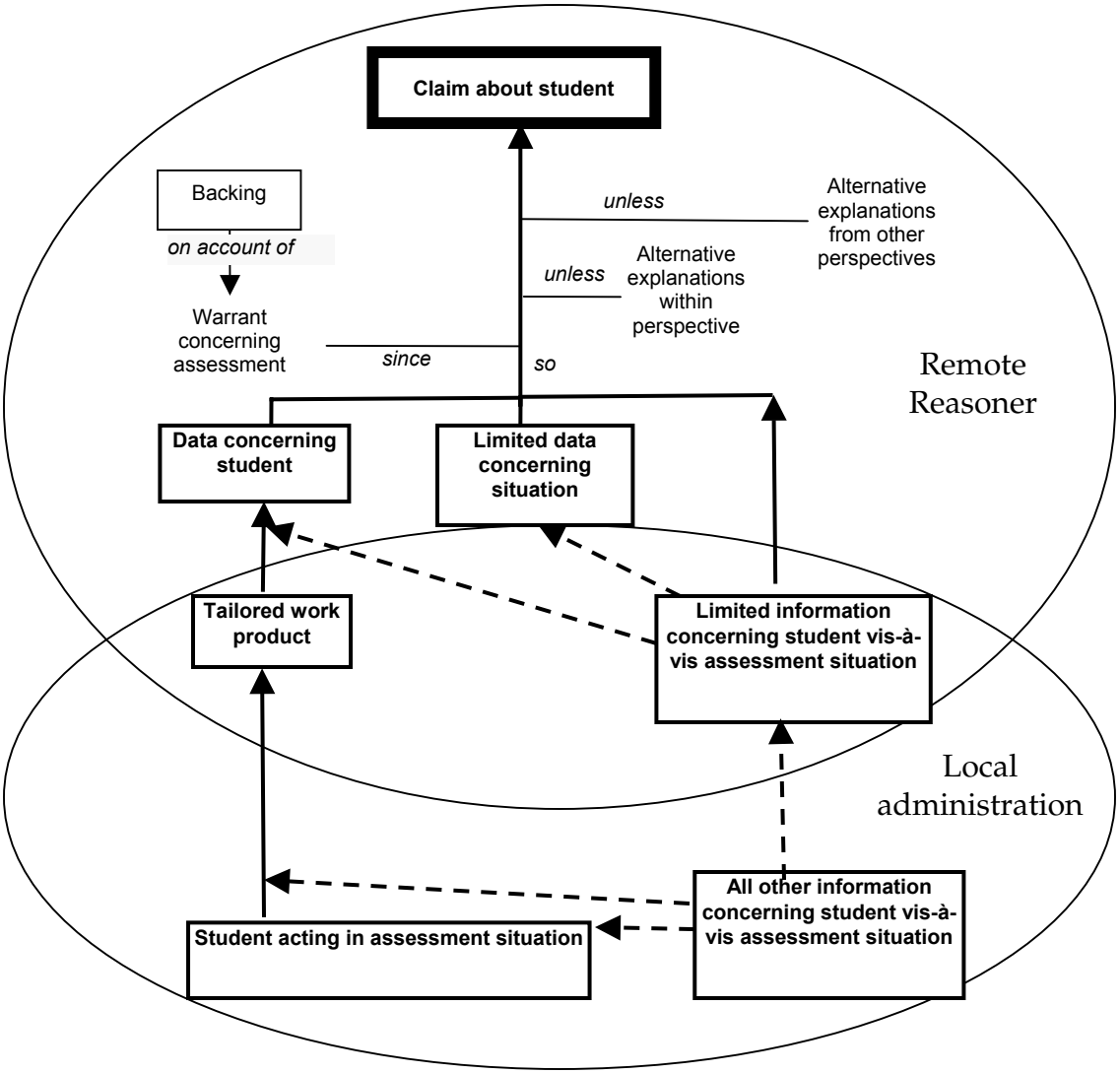


Figure 7. Toulmin diagram for an assessment argument, showing a remote reasoner with locally tailored work products and assessment situation and some limited contextual information for remote interpretation. This is the case of AP Studio Art portfolio concentration sections.

only the particular claim or observation, but other claims or observations that could be entertained. If you know what the value of a variable is, you also know what it is not. Shafer (1976) defines a "frame of discernment" as all of the possible subsets of combinations of values that the variables in an inferential problem at a given point in time might take. The term "frame" emphasizes how a frame of discernment circumscribes the universe in which inference will take place. The term "discernment" emphasizes how a frame of discernment reflects purposive choices about what is important to recognize in the inferential situation, how to categorize

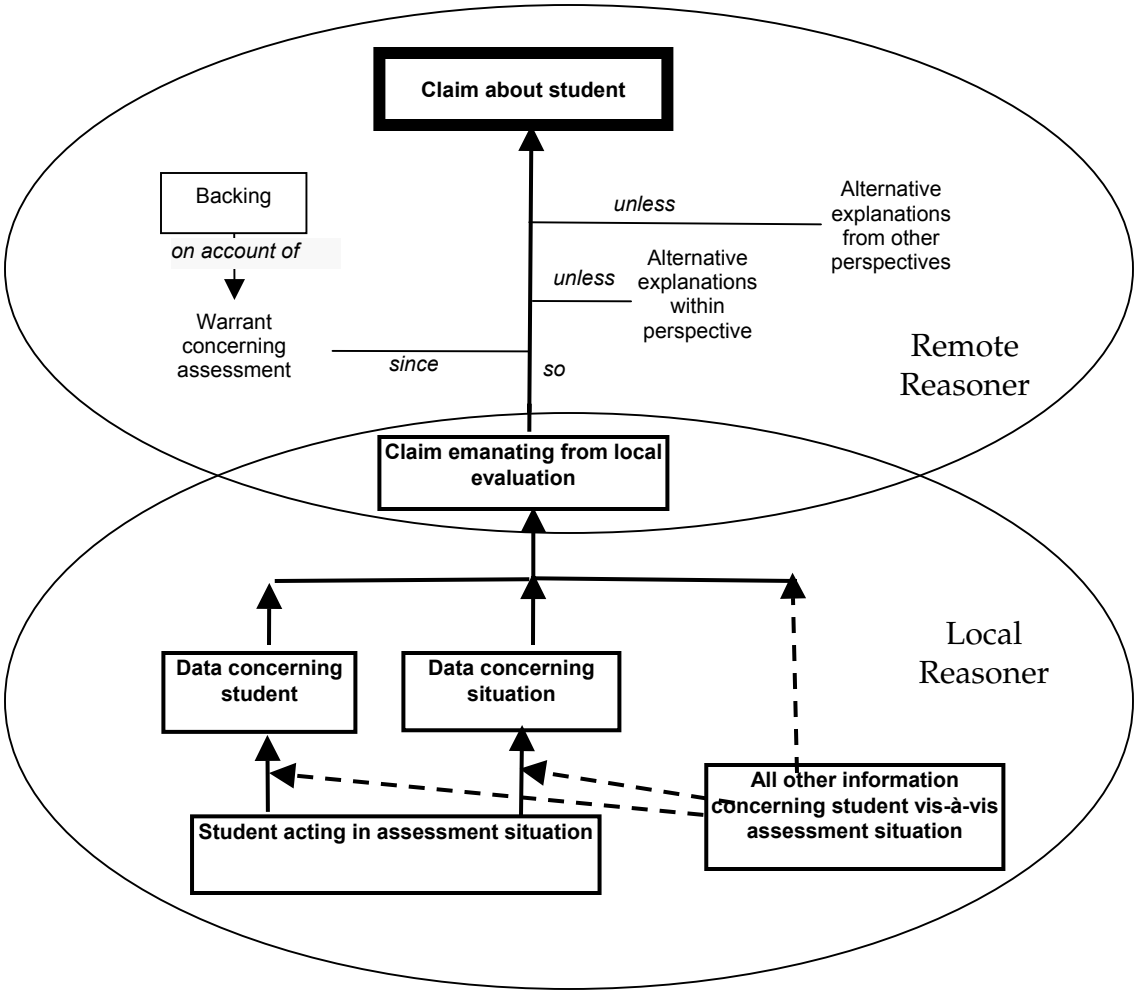


Figure 8. Toulmin diagram for an assessment argument, showing a remote reasoner obtaining summary local evaluations of students' actions within assessment situations.

observations, and from what perspective and at what level of detail variables should be defined.

The two main kinds of variables in probability models for assessment are often called student model variables and observable variables (Mislevy, Steinberg, & Almond, 2002). Both terms are a bit misleading. Observable variables are associated with aspects of students' situated actions, but they are not actually observed as such. Rather they are evaluations of things students say, do, or make in situations, through some perspective, and as has been noted above, possibly conditioned on contextual knowledge about the interrelationship between the student and the situation. AP Studio Art ratings exhibit this character: A rater maps from an emic interpretation of a body of work and a student's explanations into an etic expression

in a common framework of evaluation, a value on an observable variable (also see Schutz & Moss, 2004). Observable variables are the boundary of the probability model, and the probability model itself places no constraints on the ways, perspectives, or procedures by which values are obtained.

Similarly, student model variables should not be thought of as literal counterparts of mental capabilities or representations inside students' heads; that is, they should not be reified. Rather they represent possible ways in which students might be characterized, from the perspective in which an assessment is cast and of a nature grainsize that suits the assessment's purpose. As formal entities, student model variables may correspond to conceptions of proficiency cast in trait, behavioral, information-processing, developmental, sociocultural, or any psychological perspective. The same perspective will drive the nature of observations and the relationships between them (Mislevy, 2003)—that is, the view of proficiency and its manifestation, in the space of narratives a given probability model is constructed to support.

In a particular assessment with a preconstructed narrative space, we consider a set of aspects of skill and knowledge or propensities or exhibitions toward actions in various situations. These are the variables in a space of student models, particular configurations of values which approximate the multifarious knowledge or propensity configurations of actual students. Depending on the purpose, one might distinguish from one to hundreds of aspects of competence in a student model space. They might be expressed in terms of categories, qualitative descriptors, numbers, or some mixture of these; they might be conceived as persisting over long periods of time, or apt to change at the next problem-step. They might concern tendencies in behavior, conceptions of phenomena, available strategies, or levels or aspects of developing expertise. The particular form of the student model space in a given application is driven by a conception of the nature and acquisition of competence in the context of interest, and the goals and philosophy of the instructional component of the system.

The basic idea is this (see Mislevy, 1994, 2003, and Mislevy & Gitomer, 1996, for fuller discussion). In the narrative space there are different ways we might want to describe a student. Different things we might want to say correspond to different values of student-model variables (SMVs). Hypothetical students with different values of these variables would be likely to act differently in given situations, such as making predictions in line with impetus theory, say, as opposed to Newton's

laws, if they have certain misconceptions about force, or requiring more support to set up an investigation of an ecological problem. A model is fit from initial observations, which approximates probabilities of observable variables (i.e., salient aspects of situated actions) by hypothetical students at various possible configurations of values of SMVs. Then, in the operational assessment setting, a real student carries out actions. They are evaluated. Probabilities can be calculated to express how likely those particular actions would be from a student at any given values of SMVs. Extensions to this basic scenario include (a) being able to condition these calculations on contextual variables, student background variables, and aspects of student-situation interrelationships; (b) additional layers in models that correspond to similarities and influences of grouping variables such as schools or classrooms; and (c) effects for raters, so that variation in judgments at the level of mapping situated actions into values of observable variables can be studied. The last of these, we see below, plays an important role in AP Studio Art.

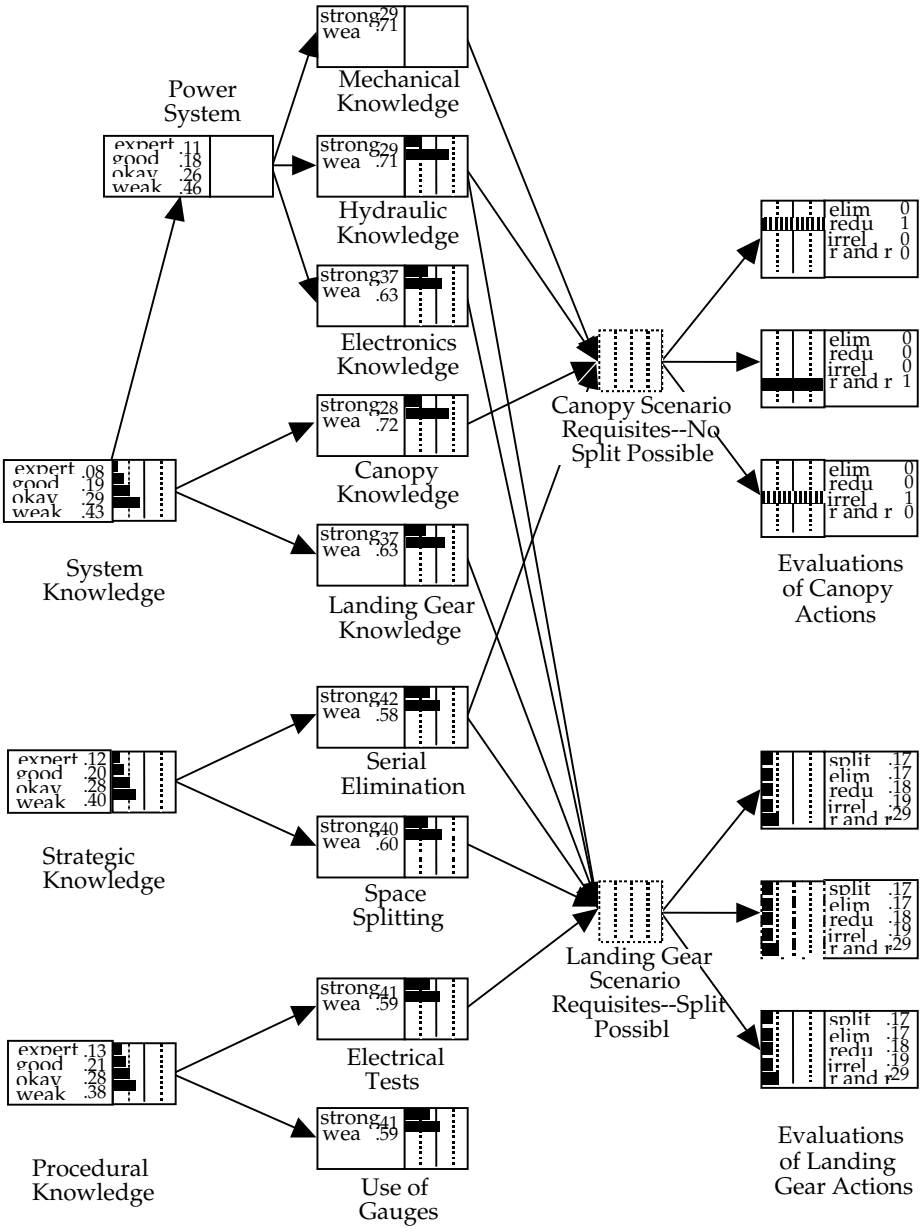
Figure 9 is the student model in HYDRIVE. Figure 9 shows a set of evidence models, or clusters of related observable variables that characterize aspects of students' actions as they work through a problem. Observable variables are defined not in terms of objective aspects of students' actions. Rather, their values take situated meaning as interpretations of sequences of actions in light of a theory of problem-solving and a history of the student's actions in the system: A student works himself into a situation; the simulator is able to define an active path of components; the simulator also computes what is knowable about the state of the system given the actions the student has taken thus far. It is then possible to characterize an action sequence as being consistent with space-splitting, serial elimination, or remove and replace strategies, or being redundant or irrelevant.

The situated meaning of the student-model variables arises from three sources:

- semantic interpretations motivated by an information-processing perspective, under which troubleshooting actions result from a conjunction of declarative, procedural, and strategic knowledge required in local contexts;

- operational interpretations arising from the way that patterns of effective and ineffective trouble shooting actions are synthesized in terms of modeled belief about higher or lower values of the student models that have been involved; and

- action-oriented interpretations in that belief shifting to low values of particular student model variables suggests a lack of skill or understanding or a type and at a grainsize that a corresponding practice or tutorial module is likely to help.



Note: Bars represent probabilities, summing to one for all the possible values of a variable.

A shaded bar extending the full width of a node represents certainty, due to having observed the value

Value of that variable; i.e., interpretations of a student's actual sequences of troubleshooting

Figure 9. HYDRIVE student model and evidence model.

In HYDRIVE, then, the student model variables are not meant to be measures of traits or simulacra of structures inside trainees' heads. They are effectively pattern recognizers that scan fairly unconstrained sequences of actions in the problem space and note incidents where practice modules are likely to improve proficiency. An information-processing perspective guided the construction of the simulator, the interface, and instructional strategy, and the context and practices of HYDRIVE's use ground the situated meaning of the student model and observable variables.

In AP Studio Art, probability-based models are used to analyze and summarize patterns in ratings across portfolios, students, sections, and raters. An AP portfolio rating session produces more than 100,000 ratings. A probability model is used to analyze information at the emic level, in the form of judges' ratings, even though those evaluations summarize individualized emic evaluations, which incorporated summaries of contextual information from the students themselves—all this in what is meant to be a common framework of evaluation, insinuated in the general rubric and fleshed out by many examples. The rater-effect statistical models that are employed originated in trait psychology, but have evolved to study patterns of variation and consistency across ratings far too numerous to examine individually in depth: Patterns such as amounts of variation expected among informed raters, signals for anomalous scores that merit further attention, and indications of the accuracy of scores obtained in a given rating design as judged against the distribution of ratings that might have occurred had all raters evaluated all work. In this way, those responsible for fairness and validity can identify atypical instances of ratings, works, or ratings, or can become aware of new styles or media that need to be accommodated into the evaluation system. In this way, tools from psychometrics are employed not to "measure traits" but to make workable a vast and geographically distributed assessment system that is grounded in the principles of situated learning.

4. Conclusion

A sociocultural/situative (SC) psychological perspective provides insights into the nature of learning and knowledge that can and should inform instruction and assessment. These insights were gained by applying detailed methods adapted from fields such as ethnography and discourse analysis. These methods are not practical to apply in their full detail for assessment on larger scales, including some within the classroom and especially ones meant to extend beyond classrooms, to

people and places distant in time and location, and for which resources are severely limited.

It is sometimes possible to design assessment practices for given purposes and contexts that are at a more apt grainsize, but which are coherent with an SC point of view at a finer grainsize. Insights and methods gained at working over the years at this grainsize in psychometrics and educational measurement can be gainfully employed in this project. One sees the variables at the coarser grainsize as emergent phenomena from the finer grainsize. The variables may even sometimes use exactly the same measurement-model machinery used by behaviorist- or trait-based assessment to synthesize evidence, but the situated meaning of the variables can be quite different. In a suitable large-scale system such as AP Studio Art portfolio assessment, one can understand values of reported variables as traces of patterns of action in situations that are locally harmonious with learning goals cast in a situative/sociocultural perspective.

References

American Council on the Teaching of Foreign Languages. (1989). *ACTFL proficiency guidelines.* Yonkers, NY: Author.

American Council on the Teaching of Foreign Languages. (1999a). *ACTFL guidelines: Reading-intermediate.* Retrieved from http://www.sil.org/lingualinks/languagelearning/otherresources/actflproficiencyguidelines/ACTFLGuidelinesReading.htm

American Council on the Teaching of Foreign Languages. (1999b). *ACTFL guidelines: Reading-advanced.* Retrieved from http://www.sil.org/lingualinks/languagelearning/otherresources/actflproficiencyguidelines/ACTFLGuidelinesReading.htm

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford, England: Oxford University Press.

Bormuth, J. R. (1970). *On the theory of achievement test items*. University of Chicago Press.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Douglas, D. *(2000) Assessing language for specific purposes.* MA: Cambridge University Press.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph Series MS-17). Princeton, NJ: Educational Testing Service.

Gasser, H. (1955). *How to draw and paint.* New York: Dell.

Gee, J. P. (2003). Opportunity to learn: A language-based perspective on assessment. *Assessment in Education: Principles, Policy & Practice, 10,* 27-46.

Gitomer, D. H., Steinberg, L. S., & Mislevy, R. J. (1995). Diagnostic assessment of trouble-shooting skill in an intelligent tutoring system. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 73-101). Hillsdale, NJ: Lawrence Erlbaum Associates.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.*), Handbook of educational psychology* (pp. 15-47). New York: Simon & Schuster Macmillan.

Greeno, J. G., Pearson, P. D., & Schoenfeld, A. H. (1997). Implications for the National Assessment of Educational Progress of research on learning and cognition. In R. Linn, R. Glaser, & G. Bohrnstedt (Eds.), *Assessment in transition: Monitoring the nation's educational progress, background studies* (pp. 151-215). Stanford, CA: The National Academy of Education.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Kadane, J. B., & Schum, D. A. (1996). *A probabilistic analysis of the Sacco and Vanzetti evidence.* New York: Wiley.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83-102.

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.

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

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

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253-282.

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2002). Design and analysis in task-based language assessment. *Language Assessment, 19,* 477-496.

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools*. New York: The Free Press.

Myford, C. M., & Mislevy, R. J. (1996). *Monitoring and improving a portfolio assessment system* (CSE Technical Report No. 402). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Resnick, L. B. (1997). Student performance portfolios. In H. J. Walberg & G. D. Haertel (Eds.), *Psychology and educational practice* (pp. 158-175). Berkeley, CA: McCutchan.

Riconscente, M., Mislevy, R. J., & Hamel, L. (in press). *An introduction to PADI task templates*. (PADI Technical Report No. 3). Menlo Park, CA: SRI International.

Salthouse, T. A. (1991). Expertise as the circumvention of human processing limitations. In K. A. Ericcson & J. Smith (Eds.), *Toward a general theory of expertise* (286-300). MA: Cambridge University Press.

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

Schutz, A., Moss, P. A., (2004). Reasonable decisions in portfolio assessment: Evaluating complex evidence of teaching. *Education Policy Analysis Archives, 12*. Retrieved April 3, 2005, from http://epaa.asu.edu/epaa/v12n33/

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton, NJ: Princeton University Press.

Steinberg, L. S., & Gitomer, D. G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science, 24,* 223-258.

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

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 Research in Education,* (Vol. 17, pp. 31-74). Washington, DC: American Educational Research Association.