Substance and structure in assessment arguments

ROBERT J. MISLEVY
Professor of Measurement, Statistics, and Evaluation, University of Maryland, College Park, MD 20742, USA

[Received on 14 May 2003; revised and accepted on 6 September 2003]

Educational assessment is reasoning from observations of what students do or make in a handful of particular circumstances, to what they know or can do more broadly. Practice has changed a great deal over the past century, in response to evolving conceptions of knowledge and its acquisition, views of schooling and its purposes, and technologies for gathering and evaluating response data. Conceptions of what constitutes assessment data, how it should be interpreted, and what kind of inferences are to be drawn can differ radically under different psychological perspectives. We see greater continuity, however, when we distinguish the structure of assessment arguments from their substance. Developments here have been more in the nature of extensions, elaborations, and refinements, as they have been prompted by changes in culture and substance.

Keywords: argument structure, assessment, evidence, psychology, validity.

Introduction

Educational assessment is reasoning from observations of what students do or make in a handful of particular circumstances, to what they know or can do more broadly. Practice has changed a great deal over the past century, in response to evolving conceptions of knowledge and its acquisition. Conceptions of what constitutes assessment data, how it should be interpreted, and what kind of inferences are to be drawn differ radically when cast under different psychological perspectives, including prominently those known as trait, behavioural, information processing, and sociocultural (Greeno et al., 1997, abbreviated GCR below; National Research Council, 2001). If we distinguish the structure of assessment arguments from their substance, however, we see greater continuity in the forms of data, their interpretation, and the inferences drawn from them. We see extensions and elaborations of recurring themes and relationships—problems of reasoning from limited numbers of observations, for example, and basing inference on the reports of imperfect raters.

The structure of educational assessments can be understood in terms of concepts and representational forms for arguments introduced by Wigmore (1937) and Toulmin (1958), broadened and extended more recently by contemporary evidence scholars such as Schum (1994), Tillers & Schum (1991), and Anderson & Twining (1991). These ideas jibe well with the conception of test validity as the grounding of the argument and the quality of the evidence for inferences based on students’ performances (Cronbach & Meehl, 1955; Kane, 1992; Messick, 1989).

This presentation begins with a brief review of Toulmin’s argument structure, including the role of claims, data, warrants, and qualifiers. We then consider how four psychological
perspectives from which assessment arguments might be cast affect the nature of claims, evidence, warrants, and qualifiers in assessment. We note the elaborations of the basic structure that are needed to accommodate increasingly sophisticated assessment arguments.¹

Toulmin’s argument structure

Toulmin (1958) provided terminology for talking about how we use substantive theories and accumulated experience to reason from particular data to a particular claim. Figure 1 outlines the structure of a simple argument. The claim is a proposition we wish to support with data. The arrow represents inference, which is justified by a warrant, a generalization that justifies the inference from the particular data to the particular claim. Theory and experience, such as empirical studies and prior research findings, provide backing for the warrant. In any particular case we reason back through the warrant, and qualify our conclusions if there are alternative explanations for the data.

In practice, an argument and its constituent claims, data, warrants, backing, and alternative explanations will be more complex than Fig. 1. An argument often consists of many propositions and data elements, involves chains of reasoning, and often contains dependences among claims and various pieces of data. Wigmore’s (1937) system of charting (modernized by Anderson & Twining, 1991) accommodates such elaborations. A further extension that is central to educational assessment is the use of statistical models as one aspect of a warrant (Mislevy, 1994).

The history of test theory in the twentieth century is a steady march toward an explication of its foundations in evidentiary reasoning, starting from a collection of practically useful testing techniques that com mingled notions of psychology, method, and purpose. By 1961, Harold Gulliksen, speaking at the 25th anniversary of the Psychometric

¹ On the role of argument structures in assessment more generally, see Mislevy et al. (2003) and Mislevy et al. (2003).
Society, was able to describe ‘the central problem of test theory’ as ‘the relation between the ability of the individual and his [or her] observed score on the test’ (Gulliksen, 1961). At the 50th anniversary meeting, Charles Lewis observed that ‘much of the recent progress in test theory has been made by treating the study of the relationship between responses to a set of test items and a hypothesized trait (or traits) of an individual as a problem of statistical inference’ (Lewis, 1986). In his influential chapter on validity in the Third edition of *Educational Measurement*, Messick (1989) described this most central concept in measurement as ‘an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment’ (p. 13).

What are the essential elements of an assessment argument?

A construct-centred approach [to assessment design] 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 behaviours or performances should reveal those constructs, and what tasks or situations should elicit those behaviours? 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. Messick, 1994, p. 16.

Note first the focus in Messick’s quotation on structure rather than substance. We will identify these central elements of assessment design with elements of Toulmin’s argument structures. Note secondly the emphasis on assessment design rather than on analysis. In assessment, the designer has the opportunity, indeed the responsibility, to fashion an assessment argument, determine the conditions of performance, indicate the kinds of observations that will be made, and lay out the procedures by which performances will be evaluated—all before the assessment is administered. The data that will be gathered and the way in which they will be used to support claims is shaped prospectively by the argument itself. This stands in contrast to most evidentiary analysis in jurisprudence, where one starts with observations (and usually gathers more along the way) but constructs arguments retrospectively (Tillers & Schum, 1991)\(^2\). The emphasis is on positing, testing, and comparing alternative theories of the events in question. The comparisons of assessment arguments under different psychological perspectives are not comparisons among theories for understanding a given set of data. They are comparisons among the data one gathers and the arguments for reasoning from it, given the psychological perspective that motivates the intended use of the assessment data.

**Four psychological perspectives**

An oft-stated axiom in evidentiary reasoning is that data are not evidence until their relationship to some conjecture, some claim, is established (Schum, 1994). In educational

\(^2\) An interesting exception is ‘sting’ operations (perhaps an appropriate analogue to educational assessments from the student’s point of view). Knowing the pertinent points of law and admissibility, an agency can construct an argument prospectively, and arrange the conditions of the sting and the actions of the agents to provide the target an opportunity to perform in a way that will likely lead to conviction.
assessment, one’s belief about the nature and acquisition of knowledge shapes the why and the what of evidentiary reasoning. Greeno et al. (1997) outline four perspectives on knowledge and learning under which instruction and assessment might be cast:

- **A trait** perspective. Messick (1989, p. 15) 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.’ Hypothetical (hence, inherently unobservable) numbers are proposed to locate people along continua of mental characteristics, just as their heights and weights locate them along continua of physical characteristics.

- **A behaviourist** perspective. The focus is on targeted behaviour in a domain of relevant situations, as both the behaviour and the situation are viewed by the assessor. Knowledge is the organized accumulation of stimulus–response associations which serve as the components of skills. People learn by acquiring simple components of a skill, then acquiring more complicated units that combine or differentiate the simpler units. Domains of knowledge can be analysed in terms of the component information, skills, and procedures to be acquired.

- **An information-processing** perspective. Epitomized in Newell & Simon’s (1972) *Human Problem Solving*, the information-processing perspective examines the procedures by which people acquire, store, and use knowledge to solve problems. Strong parallels to computation and artificial intelligence appear in the use of rules, production systems, task decompositions, and means–ends analyses.

- **A sociocultural** perspective. A sociocultural perspective stresses how knowledge is conditioned and constrained by the technologies, information resources, representation systems, and social situations with which people interact. This perspective incorporates explanatory concepts that have proved useful in fields such as ethnography and sociocultural psychology to study ‘collaborative work . . . and other characteristics of interaction that are relevant to the functional success of the participants’ activities’ (GCR, p. 7).

**Psychological perspectives and assessment arguments**

What does a psychological perspective provide an assessment argument? Everything, basically; it determines the nature of every element in Toulmin’s argument structure, and the rationale that orchestrates them as a coherent argument: what kinds of things one might say concerning students (claims), what kinds of things one wants to see (data), and why the two are related in the first place (warrants). There are always at least two classes of data in an assessment argument: aspects of the circumstances in which the student is acting, over which an assessment designer has principal influence, and aspects of the student’s behaviour in the situations, over which the student has principal influence. Additional knowledge about the student’s history or relationship to the observational situation may be further required. The traditions of the psychological perspective also determine what

---

3 People who are doing research have the luxury of being able to pick which of the myriad aspects of learning they want to focus on. People who are learning, and people who are helping them learn, don’t.
counts as backing for warrants, and the kinds of alternative explanations for performance that constitute threats to the argument.

Assessment arguments under the trait perspective

Many familiar tools of assessment began to evolve at the dawn of the twentieth century under the perspective of trait psychology, initially in a quest to ‘measure people’s intelligence.’ The essence of mental measurement under trait psychology is identifying ‘traits’ with tendencies to behave in prescribed ways in prescribed situations.

From the presumption that a given trait influences behaviour over a wide variety of situations, it follows that observations over a wide range of situations can provide evidence about that trait. In fact, writing in the context of measuring intelligence (‘g’ in his notion), Spearman posited his ‘theorem of indifference of the indicator’: ‘This means that, for the purpose of indicating the amount of g possessed by a person, any test will do just as well as any other, provided only that its correlation with g is equally high’ (Spearman, 1927, pp. 197–198).

Pet Shop Display (Fig. 2) is an example of an ‘analytical reasoning’ task, an item type used in the Law School Admissions Test (LSAT). The description of these items from the LSAT’s web site clearly takes a trait perspective: ‘Analytical reasoning items are designed to measure the ability to understand a structure of relationships and to draw conclusions about the structure.’ Such items are included in the LSAT not because lawyers or law students have to solve problems just like these in their jobs or their studies, but because studies show that students who can solve these kinds of puzzles tend to perform better in law school than students who don’t. In Toulmin’s terms, this is backing for a warrant, in this case a warrant cast in terms of a trait labelled analytical reasoning: the higher a student’s level of analytic reasoning, the more likely the student is to provide correct answers to tasks like these.

Figure 3 is the structure of the argument that leads from observing Sue give a correct answer to the Pet Shop Display to the claim that she has a high level of analytical reasoning ability. (We will address below more refined claims in terms of the values of a continuous analytic reasoning variable.) Two data elements are shown, namely the item content that satisfies the qualities stated generally in the definition of analytic reasoning and her response in that situation. There are issues of control and sequence here. The assessor was responsible for the first kind of data when the item was presented to Sue. She could respond either correctly or incorrectly to provide the second kind of data, and the basic structure of the argument would be complete. Note that the data elements and the claim are specifically about Sue, while the warrant is a broader generalization that presumably justifies inference about Sue as an instance thereof.

The item content and student performance data elements in Fig. 3 should be modelled in greater detail, to reflect an important feature of assessment: actually neither the situation nor the performance in and of themselves directly constitute the data for the argument, but rather salient aspects of them, as they are perceived by the assessor. These determinations are made in light of the purpose of the assessment and through a perspective on knowledge. A German chemistry major’s English-language paragraph on combustion might be

---

4 http://www.lsac.org/qod/questions/analytical.htm (downloaded February 26, 2003)
Pet Shop Display

Arturo is planning the parakeet display for his pet shop. He has five parakeets, Alice, Bob, Carla, Diwakar, and Etria. Each is a different color; not necessarily in the same order, they are white, speckled, green, blue, and yellow. Arturo has two cages. The top cage holds three birds, and the bottom cage holds two. The display must meet the following additional conditions:

- Alice is in the bottom cage.
- Bob is in the top cage and is not speckled.
- Carla cannot be in the same cage as the blue parakeet.
- Etria is green.
- The green parakeet and the speckled parakeet are in the same cage.

1. If Carla is in the top cage, which of the following must be true?
   a) The green parakeet is in the bottom cage.
   b) The speckled parakeet is in the bottom cage.
   c) Diwakar is in the top cage.
   d) Diwakar is in the bottom cage.
   e) The blue parakeet is in the top cage.

---

Fig. 2. An analytical reasoning item. A typical analytical reasoning item begins with a description of a situation with interrelated entities, properties, and relationships. One or more questions are posed that ask about further properties of the situation that are implied by the initial conditions.

---

evaluated for language control in English class, ignoring the chemistry context and content, but evaluated for scientific accuracy in Chemistry class, ignoring the mechanics of the language. Such considerations reflect warrants for reasoning from unique performances and performance situations to the data for the core assessment argument, as shown in Fig. 4.5 Whenever humans make determinations such as these, questions of sensitivity and objectivity appear just as they do in witness testimony in jurisprudence (Schum, 1994, p. 101 ff.). These considerations introduce alternative explanations at this stage in the full assessment argument for apparent high or low performance. Much effort in educational measurement has gone into both statistical methods and support mechanisms for monitoring and improving the evaluation of performances.

Because no single performance provides conclusive evidence about what a student knows and can do as more generally construed, most educational assessments consist of multiple observations. Figs 5 and 6 shows two ways of depicting an argument with more than one observation. Figure 5 suits a test comprised of several analytical reasoning items, each differing in its particulars but all following the same general form and requiring the same kind of reasoning. A single warrant is shown encompassing all of them. Figure 6 suits a situation Spearman described: inference about Sue’s analytical reasoning ability

5 In this figure and those that follow, we may abbreviate or omit from a given Toulmin diagram those elements that are not central to the point that is being made.
from diverse forms of evidence, such as her Pet Shop response, a teacher recommendation, and a grade in algebra class. The justification for each requires its own warrant.

Whenever there are multiple pieces of evidence, often in conflict, sometimes overlapping, the challenge becomes synthesizing their import into a final conclusion. Developments in probability-based reasoning since the 1980s have provide a solution, in the form of Bayesian inference networks (e.g. Jensen, 1996; Edwards, 1998). Claims and data become variables in the network, and qualitative warrants are the starting point for quantitative expressions of relationships between claims and data.

Psychometric models are special cases of this kind of reasoning. Although the role of probability-based reasoning in assessment is not the focus of this presentation, a few words on the topic are in order (see Mislevy, 1994, for more extended discussions). In trait-based applications, traits are unobservable variables that characterize students. Aspects of students’ responses are observable variables, which are modelled as depending in probability on the student variables. This is an expression of the warrant in a deductive direction, expectations for what observables might be if student variables were known to take any particular value. A claim is now expressed as a probability distribution that represents degree-of-belief about the values of the student variables.

Figure 7 illustrates the probability model for the similar-tasks example. Once such a model is fit and parameters have been estimated from initial data (pretest data, or ‘calibration’ data), Bayes’ theorem can be used to update belief about student variables in the light of task performances. The probability model has become an additional aspect...
of a compound warrant, which permits quantitative expression of belief and the calculus of probability to synthesize multiple, possibly conflicting, possibly overlapping, pieces of evidence. These advantages do not come for free. Additional backing is required for the probability-based aspects of the warrant, in the form of the pretest data. Additional alternative explanations for good or poor performance are introduced in connection with model misspecification and data errors.

The contemporary view of test validation concerns examining the support for and potential threats to inferences based on assessment data. Alternative explanations that arise in trait-based assessments address the scope of the trait in question. Does performance
in the assessment tasks fail to show the hypothesized relationships with some students’ performances due to measurement error? That is, would another try at a different but equivalent set of items point in the other direction? This is an alternative explanation from within the trait perspective. Might some students be solving, say, purported spatial reasoning tasks using nonspatial strategies (French, 1965)? This is an alternative explanation associated with the information-processing perspective. Do the relationships hold for some examinees but not others, as when recent immigrants were deemed unintelligent when their low scores on IQ tests could be explained by lack of familiarity with their new home (Gould, 1981, Chap. 5)? This is an alternative explanation associated with the sociocultural perspective.

As useful as trait-based assessment scores might be for the purposes of selection, classification, certification, or programme evaluation, it was ultimately their limitations for the purpose of guiding instruction that led to the rise of assessment from alternative psychological perspectives.

Assessment arguments under the behavioural perspective

Trait-based tests are not especially helpful in gauging or guiding students’ learning. As Stake (1991) points out, ‘The teacher sees education in terms of mastery of specific knowledge and sophistication in the performance of specific tasks, not in terms of literacy or the many psychological traits commonly defined by our tests.’ The behaviourist
C: Sue has a high value of Analytical Reasoning.

A1: [Alternatives re logic puzzles]

A2: [Alternatives re recommendations]

W1: [Warrant re logic puzzles]

Wn: [Warrant re recommendations]

D1: Sue’s answer to Item 1

D2: Structure & content of Pet Shop item

Dn1: Teacher recommendation about Sue

Dn2: Conditions of observation for recommendation

Fig. 6. Elaborated Toulmin diagram for multiple pieces of evidence of different kinds about Analytic Reasoning. Different warrants are needed to justify each different kind of data as evidence about Analytical Reasoning.

Fig. 7. An acyclic direct graph for a statistical model for a test composed of multiple, conditionally independent, Analytical Reasoning items. Student’s value of the unobservable Analytic Reasoning variable is denoted by $\theta$; response to Item $j$ is denoted by $X_j$, 1 if right and 0 if wrong; $p(\theta)$ is a distribution expressing what is known about $\theta$ before item responses are observed; $p(X_j | \theta)$ is a conditional probability distribution for the response to Item $j$ given any particular value of $\theta$. An updated probability distribution $p(\theta | x_1, \ldots, x_n)$ expressing what is known about $\theta$ after observing a student’s responses is obtained via Bayes’ theorem.
perspective that John Watson introduced in the early 1900s and which remained influential into the 1960s in both theory and instructional practice offers this view of increasing students’ capabilities:

The educational process consists of providing a series of environments that permit the student to learn new behaviours or modify or eliminate existing behaviours and to practice these behaviours to the point that he displays them at some reasonably satisfactory level of competence and regularity under appropriate circumstances. . . . The evaluation of the success of instruction and of the student’s learning becomes a matter of placing the student in a sample of situations in which the different learned behaviours may appropriately occur and noting the frequency and accuracy with which they do occur.


A Toulmin diagram for behaviourist assessments has the same structure as a diagram for trait-based assessments (Fig. 8), but differs as to the character of warrants, claims, and data. The substantive portion of warrants is stimulus/response linkages. Claims concern propensity for the target behaviour. One class of data concerns the features of targeted situations and the other concerns the features of performances in those situations, where the salient features of both are specified in the warrant and defined strictly from the point of view of the assessor. In contrast to Spearman’s indifference to the particulars of the situations and behaviours that constituted evidence about a trait, careful attention is now focused on specifying situations in behaviourist assessment—because propensity to the targeted behaviour in those situations directly defines the characteristic of interest about students. To draw an inference about a student’s likely behaviour in a domain of such situations, one observes the student’s actual behaviour in a sample of them. The statistical part of warrant, laid over the substantive part, is a model for success in independent trials conditional on a parameter for a student’s probability of producing the targeted behaviour.

Two kinds of tests appeared to support education from this perspective. The first, large-scale achievement testing, arose in the 1930s and 1940s to provide measures of relative proficiency in the subjects of school learning, sampled over very broad domains such as science, reading, or mathematics at a given grade level. Covering such a wide span of knowledge and skill in half an hour of testing obviously requires thin sampling, so the results of these tests are not focused enough to guide individual students’ instruction. They are meant rather to provide comparable information to determine how well students perform, compared with their grade-level peers in the sampled domains. The second type of test, criterion-referenced tests (CRTs), was introduced in the 1960s as a way of providing instructionally relevant test results to teachers (Glaser, 1963). CRTs address domains defined more narrowly in terms of explicit behavioural objectives. CRTs are designed to estimate students’ probabilities of success in a domain, with the goal of determining whether a student has ‘mastered’ it.

Alternative explanations under behaviourist assessment are fairly straightforward because the linkage among situations, performances, and claims is so direct. They include over- or under-estimating a student’s propensity toward the targeted behaviour due to incomplete or biased creation of tasks to operationally define the domain and inadequate
The important issues that are more pertinent to learning come after the assessment argument itself, regarding the use of these estimated behavioural tendencies. Does a set of propensities toward behaviour in domains defined from the assessor’s point of view adequately characterize what we want students to know, and does it support our efforts to help them learn it? The answer to these questions, emerging from the so-called ‘cognitive revolution’ in psychology starting in the 1960s, is no.

Assessment arguments under the information processing perspective

Like behavioural psychologists, cognitive psychologists who are interested in learning attend to the features of situations in which knowledge is acquired, and the contexts in which people use it. Analysis and decomposition of features of situations may again be employed. But the information-processing view goes further by taking internal processes and representations of the situation and the behaviour as targets of inference. Studies of developing expertise in domains as diverse as chess (de Groot, 1965), radiology (Lesgold et al., 1981), and volleyball (Allard & Starkes, 1980) reveal variations on a common theme. In each case, experts

... (a) provide coherent explanations based on underlying principles rather than descriptions of superficial features or single statements of fact, (b) generate a plan for solution that is guided by an adequate representation of
the problem situation and possible procedures and outcomes, (c) implement solution strategies that reflect relevant goals and subgoals, and (d) monitor their actions and flexibly adjust their approach based on performance feedback.


The claims of interest in assessment designed from an information-processing perspective, then, are not merely patterns of students’ behaviour in situations with features that are salient from the assessor’s point of view. Rather, claims concern knowledge structures, mappings of situations into knowledge structures, and performance in situations as it is mediated by those knowledge structures. Patterns of actions in suitably defined task situations still provide evidence about behavioural propensities, but this is now only an intermediate stage in an assessment argument about a student’s cognition. The inferential problem centres on students’ underlying concepts, relationships, and strategies for tackling problems in the domain, rather than solely on observable behaviours and features of problems as experts see them.

Brown & Burton (1978), for example, analysed children’s subtraction in terms of the set of so-called production rules—some correct, perhaps some ‘buggy’—that students could bring to bear on problems. Claims were cast in terms of production rules hypothesized to govern a student’s solutions. The warrant was in terms of the responses—some correct, sometimes for the wrong reasons, some incorrect, with answers that reflected buggy rules—that would be likely be produced by a student with a given set of production rules (Fig. 9). Tasks can still be grouped by features that are similar from the assessor’s point of view, but the target of inference is the student’s way of thinking that makes them similar from her point of view. A Toulmin diagram that corresponds to Brown and Burton’s assessment argument is shown as Fig. 10. The lower part of the diagram is much like that of a number of assessments cast under a behaviourist perspective: data consist of aspects of students’ actions and features of situations, arising from some propensities to such behaviour. The higher level, however, is a (possibly multifaceted) claim about the knowledge representations through which the student has perceived the situations—possibly quite different from the assessor’s—and the procedures and strategies the student brings to bear on problems as she perceives them.

Sparked by Carroll’s (1976) pioneering studies, an active area of research in assessment is exploiting what can be learned from information-processing analyses of tasks in several ways. Warrants are cast in information-processing terms. Student performances are evaluated in terms of behaviours suggested by the theory of the domain. Tasks are designed around features suggested by the theory of the domain, so that Item writers can manipulate features to increase or decrease their loads on working memory, representational form, or contextual knowledge (e.g. Embretson, 1998), and psychometrians can model tasks’ operating characteristics in terms of their cognitively salient features (Junker, 1999). Newstead et al. (2002) propose such a model for analytic reasoning items like Pet Shop, where the salient features include the number and type of constraints, and how many configurations would be compatible with the stated requirements.

Another aspect of competence that has emerged from studies of expertise is the iterative character of problem solving. Both scientists engaged in inquiry and mechanics fixing hydraulics systems, for example, generate hypotheses and provisional models, take
FIG. 9. Responses consistent with the ‘subtract smaller from larger’ bug. When the ‘subtract smaller from larger’ bug is present in a student’s configuration of production rules, problems requiring borrowing will show the characteristic pattern of incorrect responses that results from simply subtracting whichever number in a column is smaller from whichever is larger. When borrowing is not required, this bug does not affect responses; they will be correct or incorrect in whatever ways are consistent with the student’s other rules.

FIG. 10. Elaborated Toulmin diagram for inference about cognitive model in the domain of whole number subtraction. Responses and performance situations can be identical to collections of those used in a series of behavioural assessments about performance in categories of items. However, the ultimate claim is characterization of a student in terms of recognizing structures of problems and having skills and strategies to apply to solve them. Behaviour across patterns of problems of different classes is evidence for an underlying set of rules that characterize the student.
actions to test them, and revise their understanding to proceed to the next step (White & Frederiksen, 2000). This is a modelling challenge in assessment because the information from different time points is serially independent. At each time point, the performance situation now incorporates the examinee’s previous actions and their effects on the system (Fig. 11). Testing the same valve in a hydraulics problem suggests expert-level space-splitting if performed early, but it is redundant if a previous test has already eliminated that part of the system as the source of the fault.

*Assessment arguments under the sociocultural perspective*

Acknowledging the contributions of information-processing perspective, one may still question the genesis of the models, representations, and processes it addresses. Social
organizations such as families, classrooms, professions, and so on, influence the processes of acquiring, storing, representing, understanding, and creating knowledge. Many of these influences are channelled by particular ways of communicating; knowledge representations, genres, conventions, and so on. From the sociocultural point of view, ‘Learning by a group or individual involves becoming attuned to the constraints and affordances of material and social systems with which they interact’ (GCR, p. 17).

The sociocultural perspective thus proposes a view on the nature of knowledge and learning, and consequently on the nature of warrants, claims, and data for assessment framed under its aegis. In particular, ‘The situated view of assessment emphasizes questions about the quality of student 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’ (GCR, p. 37). Compared to the information-processing perspective, there is a greater emphasis on patterns of interactions of students with people and social artifacts.

The intimate connection between features of situations for acquiring and using knowledge on the one hand, and features of situations necessary for obtaining evidence about that knowledge, adds a layer of complexity to assessment under the sociocultural perspective. Contextualizing assessment decreases the assessor’s control over the features of the observational situation. It increases the burden on arranging for and identifying the salient features of both performances and performance situations. It introduces alternative explanations for good and poor performance, in connection with characteristics of the situations, people, and materials with whom the assessed student interacts. The challenges are not insurmountable. Work by Wiggins (1998) and White & Frederiksen (2000) on designing assessment to produce the kinds of learning that are valued under the sociocultural perspective is grounded in sound evidentiary reasoning. The two following examples illustrate key issues that arise in assessment from a sociocultural perspective.

The advanced placement studio art portfolio assessment. The purpose of the College Entrance Examination Board’s Advanced Placement (AP) Studio Art portfolio assessment is to determine whether high school students exhibit knowledge and skills commensurate with first-year post-secondary art courses (Mitchell, 1992). Students develop works for their portfolios in their local classes throughout the school year, through which they demonstrate the knowledge and skills described in the AP Studio Art materials. The portfolios are rated centrally by artist/educators at the end of the year, using standards set in general terms and monitored by the AP Art advisory committee. These standards are rendered in language sufficiently general to apply to a wide range of subjects, styles, and media. Their meaning is constructed over time and across sites through shared examples. Coming to learn this language, this artist’s way of seeing the world, in evaluating their own work and that of others is in fact a key learning goal of the programme. Assessment here concerns ‘questions 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 et al., 1991, p. 51).

Section B of the portfolio, the student’s ‘concentration’, is of particular interest in regard to constructing a warrant to reason from a student’s work to a rating. A concentration consists of up to 20 slides, a film, or a videotape illustrating extended work on a student-
selected theme, and a narrative describing the student’s goals, intentions, influences, and other factors that help explain the work. In the narrative, the student makes a case for how the common rubric for evaluating concentrations should be applied to her particular work. Thus the application of the warrant is effectively negotiated between the student and the assessor (Fig. 12).

A psychometric model is used in AP Studio Art to combine the scores that several raters assign to the several sections of a student’s portfolio. The student variable is not interpreted as a trait. Rather, the resulting score is viewed as a synthesis of informed and socially moderated judgments about a student’s specific accomplishments. The psychometric model is not a tool for measuring qualities of students, but a tool for managing episodes of evaluation across many portfolios, many raters, and many portfolio sections.

*Conversational competence.* The most familiar form of large-scale language testing addresses the knowledge of language per se, exercising points of vocabulary, syntax, and comprehension with discrete and largely decontextualized test items. Assessments so constructed fit comfortably into trait, and sometimes behavioural, perspectives on learning, and support assessment arguments framed in their terms. But this kind of knowledge is not
enough to use a language to achieve ends in social situations. In addition to grammatical competence, we must be concerned with the social context of language use, pragmatic considerations in using language to achieve goals, and familiarity with forms, customs, and standards of communication above the level of sentences. Conversational competence (Widdowson, 1978), for example, means being able to participate in an interaction with another person in using all of these kinds of knowledge to construct a joint understanding of a situation of mutual interest, to achieve some purpose.

Obtaining direct evidence about conversational competence requires observing a student engaging in the interaction that characterizes conversation. Four factors immediately impact the assessment argument. First, the events in conversations are not conditionally independent given the participants’ conversational competence, but are serially dependent. In addition to whatever global context a conversation occurs in, the local context for every utterance depends on what the speaker and the interlocutors have already said. Second, more than one person is acting, and each person’s actions influence and set the performance context for the others. Figure 13 shows these data relationships in the assessment argument. Third, the rules by which students’ performances are evaluated must take these dependences and interactions into account. The dependences are similar to the ones in dynamic problem-solving discussed above in connection with the information-processing perspective, but we have now added the complexities that arise when two people use interpersonal and cultural knowledge as well as domain and language knowledge, to jointly achieve some goal.

The fourth factor is that an assessment designer working under the sociocultural perspective has less control over the contextual data if the salient features of a performance situation can only emerge from interactions among individuals. As an assessment task, a minimally constrained conversation between two students can have the advantage of meaningfulness to the participants. This mitigates alternative explanations of poor performance that stem from lack of background knowledge or motivation. But the need to, say, switch from an informal to a formal register may not arise, and the observation provides no evidence about this facet of ability. On the other hand, a trained interviewer can guide a conversation in a way that provokes register-switching, facets of language, or social conventions. The performance-situation features which, under the conversational-competence warrant, are needed to obtain evidence about targeted facts of ability can be better assured—though now at the cost of introducing alternative explanations for poor performance based on the artificiality of the conversation.

The effective meaning of student variables in such a assessment arises from the features of the situations in which students perform and the features of their actions that are evaluated. The degree to which inferences about students based on observations in these particular settings (including the participants and contexts) will apply to other settings is a matter of similarities to other potential settings, and students’ tendencies to interact in those settings in similar or different ways. This is fundamentally the same question of the generalizability of test scores that arises under trait-based interpretations. In both cases it is ultimately an empirical question. How far will the meaning of scores generalize? Before the data are in, the trait psychologist’s prior is ‘pretty far’; the sociocultural psychologist thinks, ‘I wouldn’t bet on it’.
Conclusion

The forms and the uses of educational assessment have changed considerably over the past century, in response to changing views of the nature of knowledge—how it is acquired, how schooling should be organized to promote it, and how assessments should be designed to guide instruction. Continued changes, perhaps ultimately even more radical, are taking place today as a result of new technologies for gathering and analysing performance data. A closer look at the structure of the arguments beneath assessments that can appear very different on the surface reveals a deeper kind of stability. The stability is found in terms of the core argument of any assessment: we want to draw inferences about what students know and can do as seen from some perspective. That perspective tells us what kinds of things we need to see them do, in what kinds of situations, to ground those inferences. As psychological perspectives have evolved, we see in the history of assessment a number of elaborations, extensions, and specializations of a constant set of principles of evidentiary reasoning. We may therefore expect to find continued value
in knowledge representations such as Toulmin diagrams, Wigmore charts, and Bayesian inference networks to understand yesterday’s assessments, manage today’s, and design the assessments of tomorrow.

Acknowledgements

Presented at Conference on Inference, Culture, and Ordinary Thinking in Dispute Resolution, Benjamin N. Cardozo School of Law, Yeshiva University, New York, April 27–29, 2003. This work builds on research with Linda Steinberg and Russell Almond at Educational Testing Service on the structure of educational assessments. The influence of David Schum’s investigations of evidentiary reasoning is gratefully acknowledged. Geneva Haertel and an anonymous provided helpful comments on earlier versions. Support was provided by the Educational Research and Development Centers Programme, PR/Award Number R305B960002-01, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The PADI project is supported by the National Science Foundation under grant, REC-0129331 (PADI Implementation Grant). The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

REFERENCES


6 A particular strength of Wigmore diagrams is the ability to incorporate competing theories into a single representation. A user is encouraged to identify alternative explanations and sources of doubt at all levels of an argument. The Toulmin diagrams used more extensively in this presentation serve well to sketch out the main line of an argument. For more detailed work in validity argumentation, Wigmore diagrams might prove to be the superior representation.


JENSEN, F. V. 1996 *An Introduction to Bayesian Networks*. New York: Springer.

JUNKER, B. J. 1999 Some statistical models and computational methods that may be useful for cognitively-relevant assessment Commissioned paper prepared for the Committee on the Foundations of Assessment, National Research Council, http://www.stat.cmu.edu/~brian/nrc/cfa/


