**On the Structure of Educational Assessments**

CSE Technical Report 597

Robert J. Mislevy
CRESST / University of Maryland

May 2003

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
Los Angeles, CA 90095-1522
(310) 206-1532

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.

# ON THE STRUCTURE OF EDUCATIONAL ASSESSMENTS[1]

**Robert J. Mislevy**

**CRESST/University of Maryland**

**Linda S. Steinberg and Russell G. Almond**

**Educational Testing Service**

## Abstract

In educational assessment, we observe what students say, do, or make in a few particular circumstances, and attempt to infer what they know, can do, or have accomplished more generally. A web of inference connects the two. Some connections depend on theories and experience concerning the targeted knowledge in the domain, how it is acquired, and the circumstances under which people bring their knowledge to bear. Other connections may depend on statistical models and probability-based reasoning. Still others concern the elements and processes involved in test construction, administration, scoring, and reporting. This paper describes a framework for assessment that makes explicit the interrelationships among substantive arguments, assessment designs, and operational processes. The work was motivated by the need to develop assessments that incorporate purposes, technologies, and psychological perspectives that are not well served by familiar forms of assessments. However, the framework is equally applicable to analyzing existing assessments or designing new assessments within familiar forms.

## Introduction

An assessment is a machine for reasoning about what students know, can do, or have accomplished, based on a handful of things they say, do, or make in particular settings.

An assessment is more than this, of course. All assessments are embedded in a cultural setting and address social purposes both stated and implicit. Assessments communicate values, standards, and expectations. Some assessments are opportunities to extend learning. Others don't even look like assessments as we

[1] We are indebted to too many people to list here for enlightening discussions of topics addressed in this paper. We would like to acknowledge Lyle Bachman, Irwin Kirsch, Mary Schedl, and John Norris with regard to issues in language assessment, and, for their comments on an earlier draft, the editor Mark Wilson and two anonymous referees.

usually think of them; they look like conversations between a student and a teacher, or one student with another. What all assessments share, though, is reasoning that relates the particular things students say or do, to what they know or can do as more broadly conceived; that is, in terms that have meanings beyond the specifics of the immediate observations. The argument behind such reasoning is grounded in beliefs about the nature of knowledge in the domain in question, how we recognize it when we see it, and situations in which evidence about that knowledge might be manifest. This paper concerns relationships among (a) the claims one wants to make about students in order to serve an assessment's purpose; (b) the principles upon which this reasoning is based; and (c) the "pieces of machinery"—tasks, responses, rubrics, statistical routines, score reports, and the like—that one assembles to gather evidence to support claims about students.

Our work is motivated in large part by the challenge of designing new forms of assessment. Three decades of progress in fields that are central to assessment—cognitive psychology, measurement models, information technology, and learning in the disciplines—have had surprisingly little impact on everyday practice. We have come to believe the bottleneck is knowing how to make sense of rich data for ambitious inferences; more specifically, how to construct arguments that are often more complex than those that underlie familiar forms of assessment, then to assemble operational elements that also are often more complex than those of familiar assessments. Innovative assessment projects fail when insights from the noted fields are patched into largely unexamined assessment practices.

This paper lays out structures that underlie substantive assessment arguments, guide the design of assessment elements that incorporate the argument, and control the activities that realize its operation. It builds on our work on assessment design, which we call "evidence-centered" assessment design (ECD) to underscore the central role of evidentiary reasoning. ECD entails the development, construction, and arrangement of specialized information elements, or assessment design objects, into specifications that embody the substantive argument that underlies an assessment. Many of the ideas and elements we discuss are familiar. Our focus is on the web of interrelationships, between and within, assessments as substantive arguments and assessments as objects and processes. As a normative model, the ECD framework helps one understand the connections among an assessment's purpose, a conception of proficiency in the domain, the evidentiary argument, the design of the assessment elements, and operational processes. As a design

framework, it organizes thinking about domains, purposes, and delivery capabilities in a way that leads to coherent assessments. A practical advantage for large-scale programs is that designing assessments in a common framework from conception to implementation facilitates developing objects and processes that can be reused and that can operate with one another in different configurations.

After an overview of the ECD framework, we discuss assessment as evidentiary argument. We connect these ideas to the design framework that lays out the structure of an assessment, namely the *conceptual assessment framework* (i.e., the CAF; Mislevy, Steinberg, Breyer, Almond, & Johnson, in press), and the associated framework for the implementation of an assessment, namely a four-process model for assessment delivery systems (Almond, Steinberg, & Mislevy, 2001). The higher-level structures in these models suffice for the purposes of this report, so we won't detail the finer-grained structures within the models, or describe the software for creating and using the objects.

We will illustrate ideas with examples from language testing. 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. 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. Interest is therefore rising in Task-Based Language Assessment (TBLA), the process of evaluating communicative performances elicited "as part of goal-directed, meaning-focused language use requiring the integration of skills and knowledge" (Brindley, 1994, p. 74). The running examples in this article suggest how the ECD framework can be used to think through design issues in TBLA.

## Overview of the ECD Structures

This paper emphasizes the structure of assessment over the design process, but the two are hard to separate. Our understanding of structure developed as we attempted to design unfamiliar forms of assessments. The language and structures from our design work provide language and structures for analyzing an assessment, whether or not it was explicitly designed in accordance with that structure. The

structure of the paper parallels the movement through the stages of assessment design.

Figure 1 is an overview of stages in assessment design, reflecting the structures in the ECD design and delivery models.[2] An assessment design is intended to embody an argument that suits the assessment's purpose. The stages reflect successive refinement and reorganization of knowledge about the domain and the purposes of the assessment, from a substantive argument to the specific elements and processes required for its operation.

We'll start at the bottom of the figure, a model that deals with people's relationship to assessment as something experienced; namely, an operational assessment. We need a framework that describes the assessment experience from the perspective of those who participate in an assessment or use its information, for these elements and processes are what we must ultimately produce. This is the model for the four-process delivery system, which describes the functioning of an operational assessment. Let's move around the delivery-system diagram, beginning with the upper right corner:

- Presentation: Something is presented, an interaction takes place, and a response is captured.
- Response Scoring: The response is evaluated for what is possible and important to observe.
- Summary Scoring: The information in what is observed is synthesized to ground beliefs about what someone is likely to know or be able to do, and then
- Activity Selection: A decision is made about what might be useful to do next.

[2] The design process is not as linear as the figure might suggest. Iterations should be expected, moving down as decisions are made and processes are specified, and moving up as new requirements or results from field tests modify the argument.

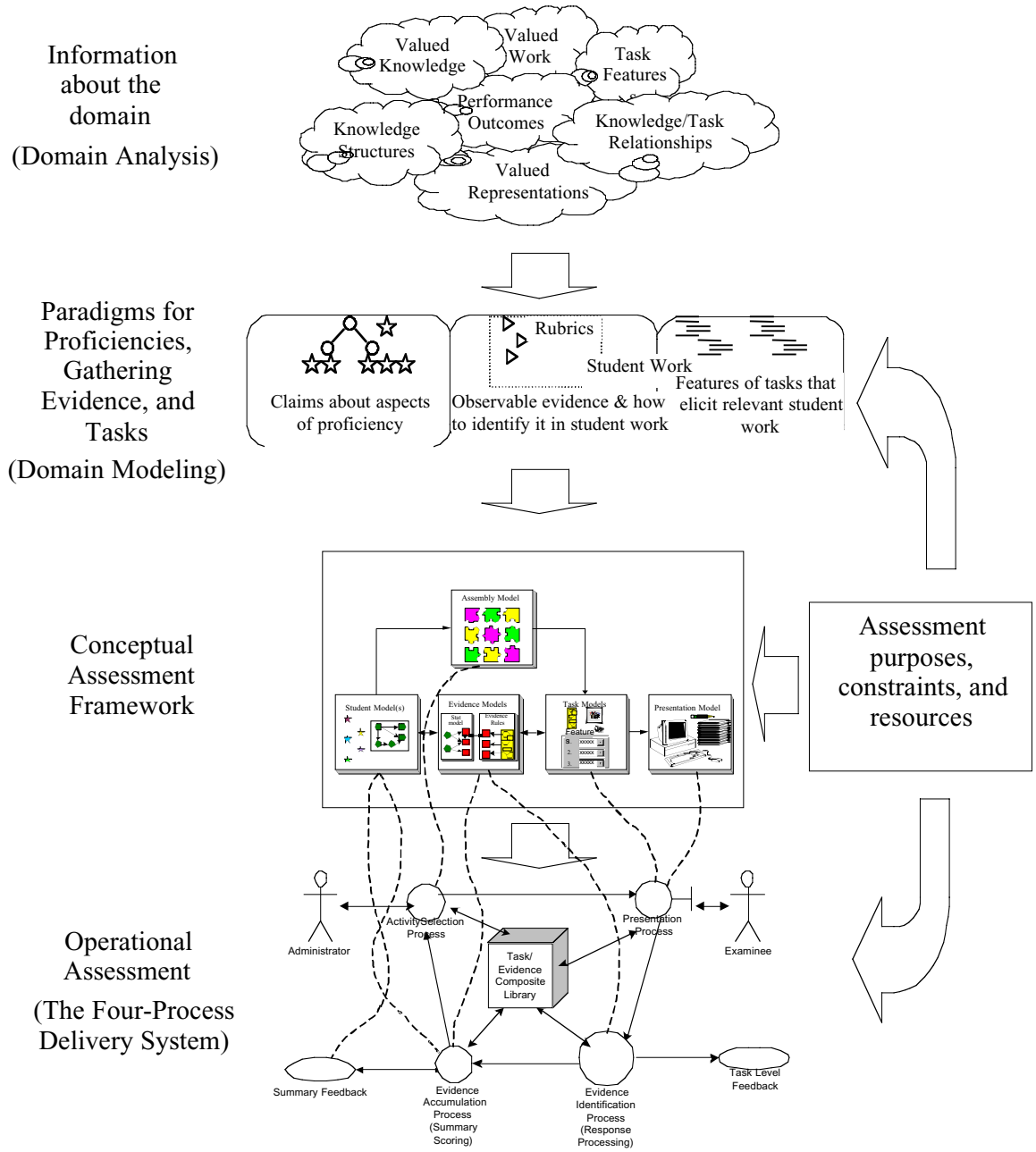



*Figure 1.* Overview of stages of assessment design.

These processes can take a great variety of forms, and they can be organized in dramatically different ways to meet different purposes. But why should we believe that any particular arrangement of tasks, responses, and scores would achieve the purposes an assessment is meant to serve? We must explore the deeper structures of the substantive argument that underlies the assessment and the design structures that connect the argument with the machinery.

Returning then to the top of the diagram, the first stage in design is Domain Analysis. It concerns marshaling substantive information about the domain—bringing together knowledge from any number of sources, then beginning to organize beliefs, theories, research, subject-matter expertise, instructional materials, exemplars from other assessments, and so on. All of this information can have important implications for the assessment, but most of it was neither originally created nor organized in terms of the structures of assessment.

In the second stage, Domain Modeling, this information is organized in terms of design objects called paradigms: Structures that organize potential claims about students and aspects of proficiency they reflect (proficiency paradigms); the kinds of things students might say or do that would constitute evidence about these proficiencies (evidence paradigms); and the kinds of situations that might make it possible to obtain this evidence (task paradigms). The focus at this stage of design is the evidentiary interrelationships that are being drawn among characteristics of students, of what they say and do, and of task and real-world situations in which they act. Here one begins to rough out the structures of an assessment that will be needed to embody a substantive argument, before narrowing attention to the details of implementation for particular purposes or to meet particular operational constraints.

The next stage of design is the CAF. The CAF models lay out the blueprint for the operational elements of an assessment; the interrelationships among the models coordinate its substantive, statistical, and operational aspects. The CAF models provide the technical detail required for implementation: specifications, operational requirements, statistical models, details of rubrics, and so on. In brief, the *student model* specifies the variables in terms of which we wish to characterize students. *Task models* are schemas for ways to get data that provide evidence about them. *Evidence models* contain two components, which are links in the chain of reasoning from students' work to their knowledge and skill: The *evaluation component* of the evidence model contains procedures for extracting the salient features of student's performances in task situations—i.e., *observable variables*—and the *measurement component* contains machinery for updating beliefs about student-model variables in light of this information. These components correspond roughly to *task scoring* and *test scoring.* The *assembly model* describes the criteria by which multiple tasks are selected for gathering evidence from students, as an assessment. Finally, the

*presentation model* gives specifications for presenting tasks, managing interactions with students, and capturing their work.

The importance of the substantive relationships forged in Domain Modeling is nowhere clearer than in the definition of student-model, observable, and task-model variables in the CAF. Each CAF variable does indeed specify a distinct aspect of pieces of assessment machinery, and they can be implemented separately. But they cannot be conceived separately. The world presents us with inseparable events, particular people doing specific things in concrete situations. Patterns we conceive in this interplay are the basis of assessment design variables, and our conceptions of them can be regularized (as they are in the ECD structures) but not uncoupled. Aspects of students' proficiencies, for example, are only a means to talk about people's tendencies to act in particular ways in situations with particular features.

A conception of the knowledge or skill one wants to measure can be a useful starting point for assessment design (Messick, 1994). So can the kinds of things one wants to see students learn to do (e.g., Wiggins, 1998), or real-world situations in which we are ultimately interested (Bachman & Palmer, 1996). These points of view correspond to an emphasis on concerns that are central to the proficiency, evidence, and task paradigms. Good assessment comes not from "choosing the right one," but by synthesizing them. Creating a collection of interlocking paradigms ensures that the elements that are highlighted in the different perspectives have been thought through and integrated.

## Educational Assessment as Evidentiary Argument

Viewing assessment from the perspective of arguments is natural once one accepts that validity, the cardinal virtue of assessment, is all about "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" (Messick, 1989, p. 13; see also Cronbach, 1989; Cronbach & Meehl, 1955; Embretson, 1983, 1998; Kane, 1992; Messick, 1989, 1994; Sarbin, Taft, & Bailey, 1960). In assessment, the data are the particular things students say, do, or create in a handful of particular situations, such as essays, diagrams, marks on answer sheets, oral presentations, and utterances in conversations. Usually our interest lays not so much in these particulars, but in the clues they hold about what students know or can do as cast in more general terms. These are the *claims* we'd like to be able to make about students, on the basis of *observations* in an assessment setting. The

nature and the grainsize of assessment claims are driven by the purpose(s) of the assessment. This task of establishing the relevance of assessment data and its value as evidence depends on the chain of reasoning we construct from the evidence to the claims (Kadane & Schum, 1996; Schum, 1987, 1994).[3]

**Example: Some Purposes and Claims in Language Assessment**

An infinite variety of claims at various grainsizes, emanating from different conceptions of language proficiency, could be envisaged to suit different assessment purposes of language testing. One useful starting point in TBLA is the real-world situations in which language is used, or the set of target language uses (the "TLUs"; Bachman & Palmer, 1996) that are central to the purpose of the assessment. What are people trying to accomplish in those situations using language? What are the important features of the situations? What do the people have to know and do, beyond vocabulary and grammar? Our running example draws on Bachman and Palmer's (pp. 253-284) illustration concerning the academic writing that non-native English speakers need to do in college. The TLUs include writing proposals, term papers, and in-class essays. We will focus on reflective writing, exemplified by an assignment in which a student must integrate ideas from written material on a given topic, then write an analysis that should be substantively responsive, lexically and grammatically correct, and structurally and stylistically appropriate to the genre.

Another starting point in TBLA is research on the competences that may be required in some way and to some degree in any given language use situation. Bachman (1990) for example, proposed a hierarchical model of the competences involved in language use, including grammatical, textual, sociolinguistic[4], and illocutionary[5] competences. Identifying target TLUs makes it possible to be more specific about the nature of competences that are at issue, and hence the claims we want to address in a particular assessment. Moreover, it helps us identify key

[3] Seeing assessment as argument, and recognizing the responsibility to establish chains of reasoning, underscores how beliefs and purposes shape assessment as much as models and methods. Educational assessment is less "measuring mental faculties" than a means for the "preservation and expansion of the intersubjectivity of possible action-orienting mutual understanding," to borrow a phrase from Habermas (1971, p. 310).

[4] Sociolinguistic competence is knowing how to use language as appropriate in social contexts, such as what topics are appropriate to discuss with close friends but not casual acquaintances, the conversational patterns of ordering food in restaurants, and the conventions of writing for academic purposes, shopping lists, and business letters.

[5] Illocutionary competence is knowing how to use language to accomplish goals such as expressing ideas, getting information, making requests, and issuing commands.

features and expectations in those TLUs. The purpose of the assessment will determine just which features and expectations are central, and which are irrelevant (Messick, 1994). The purpose that Bachman and Palmer address is non-native English speakers' placement into academic writing programs: There is a "sheltered" program tailored to non-native English speakers, although students may not have the requisite skills to enter it, or they may have enough skill to take a general academic writing class. It is assumed that all the students have displayed academic reading and writing skills in their first languages, so we can focus on the extent to which they can employ these skills in English. We will consider four purposes that are all consistent with the same conception of competence and set of TLUs, but have different implications for assessment arguments and assessment designs:

- *Purpose 1: Determining whether a student from any department can place out of the academic writing sequence.* Corresponding to this yes/no decision is a broad claim that a student can acquire and synthesize ideas from college-level reading material, control language, and use the conventions of academic writing. Because the assessment is to be used across departments, the claim does not address any specialized topical knowledge.

- *Purpose 2: Determining whether a student in psychology satisfies the department's requirement for reflective writing in the field.* This purpose is a yes/no decision like the first, and corresponds to a claim that is also broad but has been focused with respect to topical knowledge; specifically, that a student can acquire and synthesize ideas from college-level reading material concerning psychology, control language, and use the conventions of academic writing about psychology.

- *Purpose 3: For students who did not place out of the academic writing sequence, providing a summary report for instructors of sheltered and remedial classes, on each student's areas of strength and weakness as to English language usage, reading proficiencies, and academic writing skills.* This purpose concerns the same TLUs and constellation of competences addressed by Purpose 1, but finer-grained claims are required to meet its objective. For illustration, we will use claims that reflect increasing levels of proficiency in academic reading and academic writing (see Table 1). The reading claims are based on Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt, and Schedl (2000). As with Purpose 1, these claims do not address topical knowledge.

- *Purpose 4: For students preparing to take the proficiency test described in Purpose 1, providing practice on the kinds of tasks they can expect to see and feedback on the scores they might receive.* The claim at issue is the same one addressed in Purpose 1, but the assessment will now be low stakes, instructional in nature, and wholly under the control of the student. We suppose that

Table 1

Some Claims Used With Language Testing Examples

Academic Reading Claims

*Reading to find information.* A student can locate and comprehend discrete pieces of information in a text, in order to answer questions, verify or repair understanding, and find relevant parts of a text for informational purposes.

*Reading for basic comprehension.* A student who can *read for basic comprehension* can understand the basic ideas in a text and form some understanding of the main idea, but may not be able to comprehend how the supporting ideas and details form a coherent whole.

*Reading to learn.* Beyond reading for basic comprehension, a student who can *read to learn* can link the arguments and details in a text to organizing frames such as cause / effect and compare / contrast, in order to organize the information in the text and understand the author's rhetorical intent.

*Reading to integrate information across texts.* A student can read to integrate information and arguments across multiple sources by generating an organizing frame that is not contained in any of those sources.

Notes:

- These claims are based on the purposes for reading discussed in greater detail in Enright et al., 2000.
- The claims are ordered in the sense that with similar sources, use situations, and topical material, a student who could *read to learn,* say, could also *read for basic comprehension* and *read to find information.*
- To be fully specified, a claim would further need to address use situations, topical material, features of texts, etc. Further, these claims could be broken down into finer levels of details to concern situations, topics, and features, as well as strategies and organizing frames a student can use in what kinds of situations.

Academic Writing Claims

*Providing discrete information.* A student can express knowledge or opinions in written English, using words, phrases, or fragments of sentences.

*Sentence-level writing.* A student can express knowledge or opinions using grammatical English sentences.

*Paragraph-level writing.* A student can express knowledge or opinions using multiple sentences that are grammatical, in formal register, and organized in terms of frames, such as cause / effect, compare / contrast, or theme with supporting details.

*Extended academic writing.* A student can express knowledge or opinions in the genre of expository academic writing. This writing is marked by the use of the formal register, nested organizing frames, and rhetorical devices for cohesion.

Notes:

- These claims presume that the student has the pertinent knowledge or opinions to work with, and the issue is expressing them in the conventions of academic written English.
- The claims are ordered in the sense that with use situations and topical knowledge, a student who could carry out paragraph-level writing, say, could also carry out sentence or discrete element writing.
- To be fully specified, a claim would further need to address use situations, subject matter, expected features of produced texts, etc. Further, these claims could be broken down into finer levels of details to concern situations, topics, and features, as well as elements of usage and structure.

tasks from previous administrations of the placement assessment are available for use, and a computer-based expert-system is available to provide estimates of the scores that raters would assign students' essays (e.g., the e-rater system described in Frase et al., in press).

## The Structure of Arguments

We connect claims and observations through a web of inference. Some links depend on our beliefs about the nature of knowledge and learning. What is important for students to know, and how do they display that knowledge? Other links depend on things we know about students from other sources. Do they have enough experience with a computer to use it to solve an interactive physics problem, or will it be so unfamiliar as to hinder their work? Links in some assessments use probabilistic models to communicate uncertainty, because different students are administered different tasks or because evaluations are obtained from raters who don't always agree.

Toulmin's (1958) schema for the structure of arguments (Figure 2) provides useful terms and representations for analyzing assessment arguments. During assessment, reasoning flows from observations to claims about students. For reasons that will become apparent, we call this *reverse* reasoning, or induction. The *claim* (C) is a proposition we wish to support with the data. In the case of assessment, this is a statement about what a student knows or can do. The simple example in Figure 3 shows a claim about Sue being able to use specifics to back up her descriptions of literary characters. The *data* (D) is a proposition that expresses our observation. In assessment, the data are something a student says or does. The Figure 3 example shows as data Sue's use of specifics about *Hamlet* in a particular essay she has written.

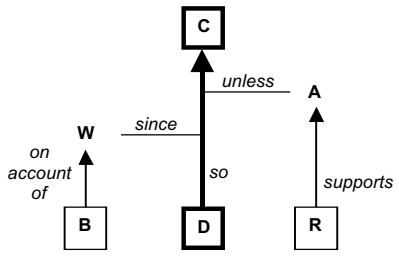



*Figure 2.* Toulmin diagram of the structure of arguments.

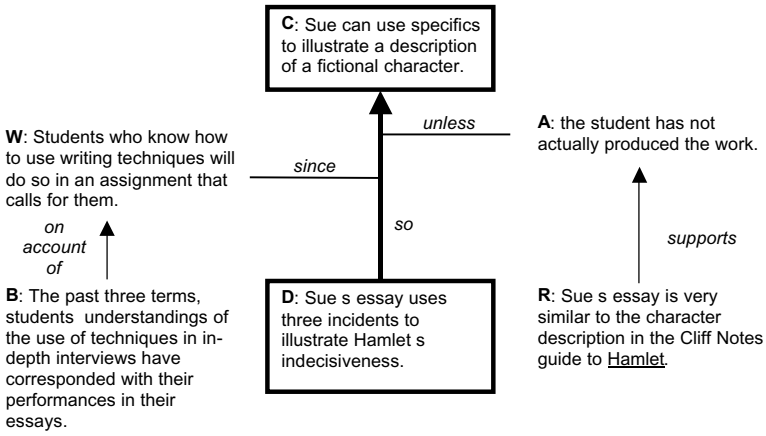



*Figure 3.* A simple assessment argument depicted as a Toulmin diagram.

The arrow represents inference, which proceeds by means of a *warrant* (W). A warrant is a generalization used to justify the inference from the particular data to a particular claim. The warrant in the example is that students who are able to support character descriptions with specifics will do so in tasks like the one at hand. Note that the warrant is cast in the opposite direction of the desired inference. Warrants express *forward* reasoning, or deduction: Under certain conditions, we expect certain outcomes (e.g., Students having the ability to use specifics usually do so in their essays).

A strong warrant such as a syllogism tells us to expect an outcome with certainty. The warrant is in that case the major premise of a logical argument, the observation is the minor premise, and the claim follows inexorably. More often, a warrant tells us only that certain outcomes are more or less likely. Knowing what outcomes to expect in general (the forward reasoning expressed by the warrant for the typical case), our inference from the specific outcome (the reverse reasoning desired for the particular case) is justified—to a degree that we are responsible for determining. Incorporating a statistical model into the warrant admits, and attempts to quantify, these possibilities.

Warrants themselves require *backing* (B), in the form of theories, research, data, or experience. The substantive foundations of warrants in assessment are our beliefs about the nature of knowledge and how it is evidenced. *Alternative* (A) or counter explanations for the observed *data* (D), may be required to qualify the inference. For example, considerations of familiarity, motivation, opportunity to learn, and conditions of testing can erode the value of students' performances as evidence

about their competence. Alternative hypotheses too may be supported or weakened by *rebuttal* (R) data. In practice, the claims, data, warrants, backing, and alternative explanations in assessment will comprise multiple propositions and data elements, involve chains of reasoning, or contain dependencies among claims and various pieces of data.

The claims in assessment are based on a conception of knowledge and skill in some domain of interest. Evidentiary reasoning *per se* is agnostic as to the nature of this conception. The challenge of reasoning about students from limited data arises whether we characterize students in strictly behavioral terms, or from a cognitive or situative psychological perspective. We will be able to use the same argument structures, even though the nature of our claims about students, what we observe, and how we interpret it, will vary accordingly.

The primary source of warrants is the same conception of knowledge and skill, now focusing on what can be seen when people put that knowledge and skill to use. When we embed the relationships among students, tasks, and performances within a statistical (measurement) model, we elaborate the substantive warrant so as to support probability-based inference. This allows us to synthesize information from disparate sources into a common framework and characterize its evidentiary value, and to reason through sometimes-complex relationships among what we observe and what we want to infer. Correspondingly, however, reasoning through a probability model opens the door to alternative explanations that concern failings of that model.

There are two main kinds of data in assessment arguments. There are the things students say, do, or produce in the assessment setting. The students themselves bring about these data. But student data are produced in particular situations, and warrants concern the relationship between students' actions and the features and conditions of the situations. The second essential kind of data in assessment, then, are the features and conditions of the situations. The assessors bring about these data. They may be determined a priori by external assessors, or negotiated among assessors and the assessed in ways that more flexibly meet the conditions of warrants, as we will discuss later.

**The Role of Context**

We have begun to discuss how students' behaviors in assessment settings and the features of those settings are the data from which we reason in assessment. This

is sufficient for assessment from a strict behaviorist perspective; the claims of interest address students' tendencies to exhibit behaviors in specified classes, in situations with specified features. But for claims motivated from cognitive, situative, or developmental psychology, the relationship between the assessment setting and the student's previous experience may also be important in the assessment argument.

Consider, for example, the reading proficiency scale developed by the American Council for the Teaching of Foreign Languages (ACTFL, 1999). Part of the description of a Mid-Intermediate level is "able to read texts that impart information to which the reader brings personal knowledge." In contrast, Advanced-Plus says "able to understand texts which treat unfamiliar topics and situations." The difference isn't what's in the task, but in how what's in the task relates to what's in the student's background. If you know the relationship for a particular student and a particular text, you can use this information as further data to sharpen your argument. If you know (or have arranged conditions so that it is the case) that a student *is* familiar with a topic, you can eliminate a possible counter-explanation for poor performance. If you know (or have arranged) that a student *is not* familiar with a topic, you must consider unfamiliarity as a possible counter-explanation for poor performance—but on the other hand, you have created a situation in which you know you can get evidence about the student's capability to use language for learning about unfamiliar topics. But if you *don't know* the relationship, you have to take into account that poor performance can be due to lack of familiarity with a text or a lack of language skills.

Not knowing the relationship between students and tasks is a source of low generalizability with performance tasks, when they are used in "drop in from the sky" assessments; little is known about the relationship between a student's experiences and task demands, expectations, topical familiarity, knowledge representations, and so on. Exactly the same response to exactly the same task can provide a lot of information to one assessor (e.g., the student's teacher) and very little to another (e.g., the chief state school officer) if only the first knows how the task is related to the student's experiences.

**The Case for Standardization**

The assessment structures we will describe can be applied to a wide range of assessment forms, from classroom quizzes and standardized achievement tests, to

coached practice systems and computerized tutoring programs, to the informal conversations students have with teachers. In all but the last of these examples, a framework has been predetermined for the kinds of data that will be gathered, the kinds of claims that will be made, and the structure of reasoning that will be used to justify the inference. In the last, decisions about kinds of observations, tentative hypotheses, and reasoning from one to the next, are unconstrained and made on the fly. We agree with Sarbin et al. (1960) that the same kinds of reasoning structures are being used either way, and we hold that the ECD perspective can be used to analyze both. Our attention in this paper is on the former.[6]

The idea is this: If we foresee that similar data will be required 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 well-designed protocol for gathering data addresses important issues in its interpretation, such as making sure the right kinds and right amounts of data are obtained, and heading off likely or pernicious alternative explanations. Only when the protocol has been violated or the specifics are atypical must the details of a given instance be explored in depth.

Large-scale assessment is made practicable in education by laying out ahead of time the argument for what data to gather and why, from each of many students who will be assessed. The particulars of the data will vary from one student to another, and so will the resulting claims. But the same kind of data will be gathered for each student; the same kind of claim will be made; and the same argument structure will be used in each instance. This strategy offers great efficiencies, but it admits the possibility of cases that do not accord with the common argument. The assessor thus has two distinct responsibilities:

- *Establishing the credentials of the evidence in the common argument.* To the extent that the same argument structure holds for all the students it will be used with, the specialization to any particular student inherits the backing that has been marshaled for the general form. Rational analyses and large-scale statistical analyses can be used to test an argument's fidelity at this macro level.

- *Detecting 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

[6] Reasoning within structures that are predetermined in this sense might be described as hypothesis-driven assessment, as opposed to hermeneutic or clinical (in the terminology of Sarbin et al., 1960).

additional data or different arguments, often on a case-by-case basis. A conscientious assessment system will minimize the frequency with which these situations occur, but draw attention to them when they do.

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. Each has different implications for the assessment argument as well as for administration. For example, few large-scale assessments are more open-ended than the Advanced Placement Studio Art portfolio assessment (Myford & Mislevy, 1995). 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.

**The Role of Probability-Based Reasoning**

Formal probability-based reasoning, in the form of measurement models, is part of the ECD structure, although it may not be needed in informal assessments. Probability-based reasoning plays an important role in assessment when we must communicate and defend our arguments, across time and space, with regard to not only *what kinds* of evidence we have marshaled for a claim but also *how much* evidence we have. When used in an assessment argument, a measurement model becomes part of the warrant. It enables more precise reverse reasoning, from multiple and diverse observations to probable values of student-model variables. This added aspect of the warrant requires backing, specifically in the form of estimating the model—i.e., pretesting, to fit the form of the model and calibrating tasks or families of structurally similar tasks (Sheehan & Mislevy, 1990). The possibility that a model is inappropriate as a whole or for particular students gives rise to possible counter-explanations for both good and poor performance.

Measurement models address the probabilities distributions of variables concerning aspects of things students say, do, or make in specific assessment settings (observable variables), as functions of variables associated with students (student-model variables) that are relevant across situations. The student-model

variables can be defined in terms of traits or competences, but they can also be defined in terms of tendencies to behave in certain ways in situations with particular features. We will discuss the relationship between *claims* and *student-model variables* in more detail later, but it is relevant to state here that student-model variables have two aspects. First, formally, they are pieces of machinery we manipulate with probability calculus to accumulate evidence across observations. Second, they accrue substantive meaning through our assessment arguments, and the interrelationships we establish among them, student behaviors, and task features.

Choosing to manage information and uncertainty with probability-based reasoning, with its numerical expressions of belief in terms of probability distributions, does not constrain one to any particular forms of evidence or psychological frameworks. That is, it says nothing about the number or nature of data, or about the character of the performances, or about the conditions under which performances are produced. And it says nothing about the number or nature of student proficiency, such as whether students are characterized in terms of numerical values in a differential psychology model, production-rule masteries in a cognitive model, or tendencies to use resources in a situative model. In particular, using probability-based reasoning does not commit us to long tests, discrete tasks, or large samples of students. For example, probability-based models have been found useful in modeling patterns of judges' ratings in the very open-ended AP Portfolio Art assessment (Myford & Mislevy, 1995), and in modeling individual students' use of production rules in a tutoring system for physics problems (Martin & VanLehn, 1995).

In a measurement model, then, a student is modeled in terms of variables $\theta$ that correspond to facets of knowledge or tendencies in behavior that suit the purpose of the assessment, and the data $X$ are values of variables that characterize aspects of observable behavior. Either or both could be vector-valued, if multiple aspects of proficiency or of performance were at issue. As a working hypothesis, we posit that student-model variables account for observable variables in the following sense: For people with a given value of $\theta$, there is a probability distribution of possible values of $X$, say $p(X \mid \theta)$. This is a mathematical expression of what we might expect to see in data, given any possible values of student-model variables—a relationship expressed in the forward-reasoning direction of a warrant. The statistical concept of *conditional independence* formalizes the working assumption that if the values of the

student-model variables were known, there would be no further information relevant to those patterns in the details.[7]

The way is now open for reverse reasoning, from observed $X$s to likely $\theta$s, through Bayes Theorem. Suppose our belief about a given student's $\theta$ is expressed by the probability distribution $p(\theta)$ prior to learning the value of $X$, and the conditional distributions $p(X \mid \theta)$ give the probability distributions for the possible values of $X$ given any particular value of $\theta$. Bayes' Theorem says the updated probability distribution expressing belief after observing $X=x$ is the posterior probability distribution $p(\theta \mid X=x) \propto p(X=x \mid \theta)\, p(\theta)$. This formulation allows for belief about $\theta$ to (a) be accumulated over many observations, (b) synthesize information from disparate forms of evidence, and (c) use different data for different students.

The most common way that evidence is accumulated over assessment tasks is summing of "item scores," produced task by task, into "test scores." Embedding this thinking in formal probability-based reasoning is the basis of classical test theory and its descendent item response theory (IRT; more about this later). This approach suits Purposes 1 and 2 in the language assessment example satisfactorily, even though they employ complex performances and evaluations. But the evidence coming from tasks in the form of observable variables can vary in number or kind from one task to another, and in which aspects of proficiency it reflects. Multiple aspects of proficiency can be addressed, informed by evidence in different combinations in different tasks. More complex relationships between $X$s and $\theta$s can be envisaged, and indeed must be to support the data gathered in more ambitious assessments. Measurement models that are up to the job will be illustrated later with examples for Purposes 3 and 4.

No matter what the substantive perspective or the form of data, an assessor can use probability-based reasoning to help meet the responsibilities one incurs when using a standard argument form. Never fully believing the statistical model we are reasoning through, we are bound to examine model fit, both overall and at the level of individual examinees. We must examine the ways and the extent to which the real data depart from the patterns in the data, calling our attention to failures of

[7] The fact that every detail of a student's responses could in principle contain information about what a student knows or how the student is thinking underscores the constructive and purposive nature of modeling. We use a model at a given grainsize or with certain kinds of variables not because we think that is the "true" model, but rather because it adequately expresses patterns in the data for the purpose of the assessment.

conditional independence—places where our simplifying assumptions miss relationships that are probably systematic, and possibly important, in the data. Finding substantial misfit causes us to re-examine the arguments that tell us what to observe and how to evaluate it.

**Sources of Information**

Experience and theory in any given sphere of human activity provide many kinds of information about the nature of knowledge in that arena, how people acquire it, and how they use it. The information has come to exist because people found it useful for working, learning, and interacting. This information is essential in assessment, but it has not been gathered or organized with assessment in mind. The assessment designer needs to acquire such information and begin to recognize within it those bits of information, patterns, structures, and relationships, from which they will begin to organize assessment arguments and sketch assessment objects.

We refer to this stage of assessment design as *domain analysis.* The following categories are useful in domain analysis for identifying recurring kinds of information that are relevant to assessment arguments. (These categories are built in to the ECD software, which provides tools for tagging, sorting, and relating them.)

- Valued Work. These are real-world situations in which we can see people doing the kinds of things and using the kinds of knowledge we care about. They can inspire tasks, provide examples of performances, and offer clues about important features of performance situations.

- Task Features. These are recurring and important features of the situations in which valued work can be observed. They provide clues for features of assessment tasks, the part of the data the assessment designer can control to some degree in order to focus evidence, [determine stress on how different aspects of knowledge, and head off alternative explanations for performance.

- Representational Forms. Becoming proficient in a domain entails learning to work with the schematic, graphic, and symbolic representation forms of the domain. In assessment, they are the mechanisms by which information is presented to examinees, and the mechanisms through which examinees respond.

- Performance Outcomes. What are the kinds of things people say, do, or make that indicate to us that they have succeeded or failed in their endeavors, or that they have employed certain knowledge? These observations are the seeds of rubrics and scoring algorithms.

- Valued Knowledge. What kinds of knowledge and skill are considered to be important in the domain? After all, aspects of proficiency are based on the observation of regularities in peoples' behavior across situations, or observed differences among people as to performance, learning, or success.

- Knowledge Structure and Relationships. There may be conceptual structures for organizing aspects of knowledge in a domain, such as curricula, knowledge maps, production-rule analyses of performance, and studies of how competence develops.

- Knowledge-Task Relationships. Expert-novice research provides examples of studies of the relationships among knowledge, performance, and situational features in a wide range of domains (Ericcson & Smith, 1991). Here we find clues about how features of tasks help elicit evidence about targeted aspects of knowledge.

**Organizing the Information in Terms of Assessment Structures**

The next stage of design, labeled Domain Modeling in Figure 1, is sketching out potential variables and substantive relationships. Its structures lay out and interweave the threads in an assessment argument in a way that presages the more technical models in the CAF. Domain modeling consists of systematic structures for describing the proficiencies that are of interest, ways for getting observations that evidence proficiency, and ways for arranging situations that afford a student the opportunity to provide the evidence (c.f. Messick, 1994, p. 16). In the ECD framework, these design structures are called proficiency, evidence, and task paradigms, respectively. Figure 4 highlights key relationships between elements of these structures and elements of the assessment argument; the warrant justifies relationships among the paradigms.

Considerations of context play out in the construction of proficiency, evidence, and task paradigms. They shape the elements, the conditions, and the processes of an assessment, in ways that may not be apparent in the operational machinery but which are essential to the underlying argument—choices for the details of operational definitions, for example, or conditions for administration, prerequisite instructional experiences, or limits and conditions of interpretation. In terms of assessment design objects, for example, a context effect may be modeled both as part of what we posit about students (in circumscribing the meaning of a variable in the student model), and as part of what we extract from their performances (in evaluating an observable variable in light of what else is known about the student or the situation).

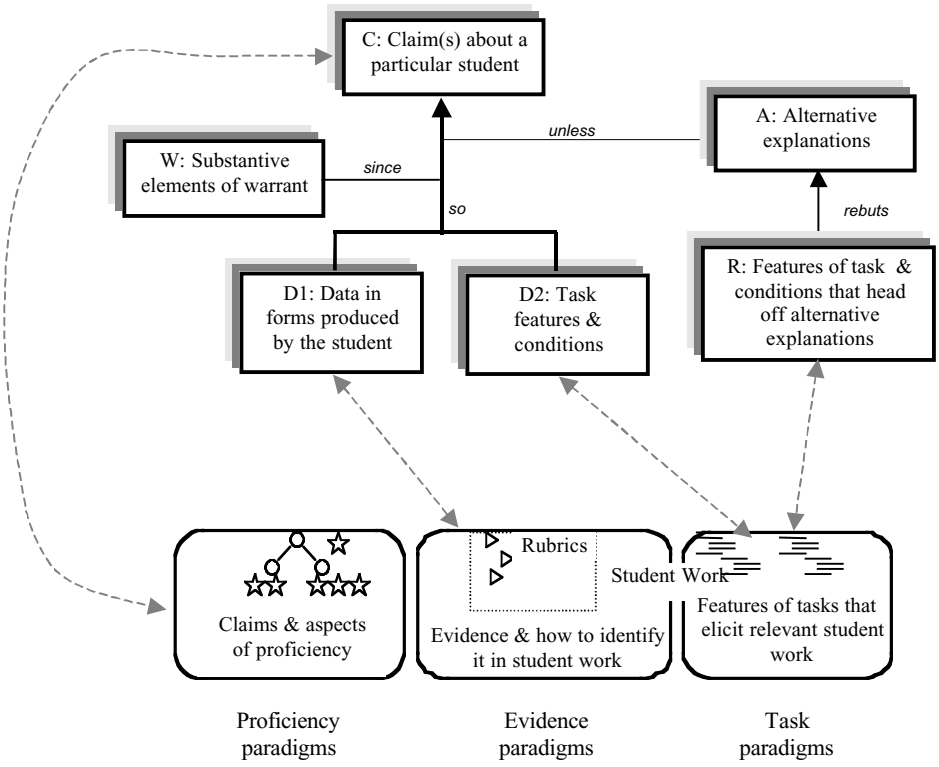



*Figure 4.* Links between the assessment argument and the paradigms in the domain model.

## Example (Cont.): Proficiency, Evidence, and Task Paradigms for Language Assessment

*Proficiency paradigms***.** In our academic writing example, we are interested in writing that requires knowledge of vocabulary and syntax, brought to bear in situations that require analysis of information within and across academic texts, expressed in the conventions of academic writing. Grammatical and textual skills are being called upon both receptively (in reading) and productively (in writing); we could identify finer-grained subcomponents of this knowledge. Sociolinguistic demands are limited to those needed to produce text in academic style. Topical knowledge is required in the motivating TLUs; some of the assessment purposes concern claims about students' topical knowledge, but others don't. This same conception of competence in academic writing will lead us to measurement models with different kinds and numbers of student-model variables in the four examples, as suit their different purposes.

*Evidence paradigms.* At this point, we are thinking broadly about the kinds of things one might see students say or do that give clues about what they know or can do. *Reading to find information,* in Table 1, for example, is marked by rapidity and accuracy, both of which can be operationally defined in many ways. Topical knowledge is evidenced when students identify or produce correct and appropriate information and relationships. *Reading to integrate across texts* is evidenced when students identify or produce organizing structures for information across multiple texts. Some observations that will provide evidence about reading to integrate may call upon other knowledge or competences as well, such as writing competence or topical knowledge. We will see later how measurement models can help sort out these natural confoundings.

Students evidence *sentence-level writing* competence (Table 1) when they identify or produce grammatical sentences. Elements of vocabulary or grammar can be addressed in isolation, as they are in discrete skills assessments, but the same kinds of observations can be made from data from more complex performances. In more complex performances, however, these observations can also depend on additional aspects of competence in ways that need to be addressed in the evidentiary argument and in the machinery that embodies it. Quality of ideas is evidenced when students identify or produce appropriate inferences about texts. *Extended academic writing* competence is evidenced when students select or produce appropriate and correct organizing structures for an issue that cuts across texts, and when they place or produce appropriate information within that structure.

*Task paradigms.* Here we address the relationships between the kinds of observations discussed in the preceding paragraph and the features of situations that enable students to display that evidence. Messick (1994) discusses how to think about the kinds of features to include or avoid in performance tasks in order to focus on the knowledge that is of interest, and minimize the impact of knowledge that is not of interest (i.e., sources of *construct relevant* and *construct irrelevant* variance). For example, to create an opportunity for students to demonstrate integration across texts, we must provide multiple texts, or must provide a text that is related to another they are familiar with. To be able to get evidence for specified levels of extended academic writing, we must define task features that characterize response requirements such as length, formality, and genre. These are examples of task features that an assessment designer can define and manage in order to provide

evidence about targeted aspects of competence, in different combinations, at different levels of stress.

## The Conceptual Assessment Framework

The CAF layer of Figure 1 shows and suggests relationships among the student model, evidence models, task models, assembly model, and presentation model. Taken together, they direct the choice and construction of pieces of operational machinery—particular tasks, rubrics, statistical models, and so on—in accordance with an assessment argument. An operational assessment will generally have one student model, which may contain many variables, but it may use several task and evidence models to provide data of different forms or with different rationales in order to get evidence about students' proficiencies in complementary ways.

We will discuss these models in an order that suits the language assessment examples. Much progress has been made in recent years in understanding the nature of language proficiency, and in the situations people use language—the substantive underpinnings for student and task models. The challenge has been connecting these insights in a way that leads to well-structured inferences. We therefore discuss student, then task, then evidence models, followed more briefly by assembly and delivery models. In practice, the models for a given assessment are not constructed sequentially but jointly, generally iteratively—because the full meaning of any model or any element within a model emerges only from its interrelationships with other elements throughout the structure. Design decisions for the various models must be coordinated from the start for an assessment to embody a complete and coherent argument.

### The Student Model

The student model (SM) is the blueprint for the part of the assessment machinery that is used to manage evidence about students. It may contain a single variable, reflecting an overall proficiency, or multiple variables, characterizing many aspects of knowledge, skill, strategies, competences, and so on. Technically, it consists of a possibly vector-valued parameter, which we will denote by $\theta$, and an accompanying joint probability distribution $p(\theta)$.

Student-model variables are the bridge between substantive claims and the statistical machinery used to accumulate evidence. Claims may be more or less detailed, and they can be expressed in terms that range from everyday language to

jargon, but in any case they are substantively meaningful statements about students. A student model is a mathematical structure containing variables that can take a range of possible values, and a joint probability distribution expressing relationships among these variables. Students' unique constellations of knowledge, proficiency, strategies, and tendencies to behave in certain ways in situations with given features will be approximated by configurations of values of SM variables. A probability distribution is used to express belief about a particular student's values in the SM variable space at any given point in time. Initially, this distribution may be quite vague. As evidence becomes available in the form of things the student says or does, we update the distribution to reflect revised beliefs. Both $\theta$ and $p(\theta)$ characterize our thinking about a student (notably, in a form we know how to update when evidence becomes available) rather than the student per se.

SM variables derive their properties as machinery from the domain of mathematical probability. Exactly the same variables and distributional forms might be used in assessments for different purposes, in different domains, or at different grades. They have no substantive meaning until we begin to use them to signify accumulating evidence of some specified type about a student. In distinction from the richness or complexity of the proficiency paradigm through which they are interpreted, the number of student-model variables in a given assessment depends on the purpose of that assessment. A single variable characterizing overall proficiency in algebra might suffice in an assessment meant only to support a pass/fail decision. A coached practice system to help students develop the same proficiency might require a finer grained student model, to monitor how a student is doing on particular aspects of skill and knowledge for which we can offer feedback.

A student model can be viewed as a fragment of a Bayesian inference network, or Bayes net (Almond & Mislevy, 1999). A Bayes net is a full probability model for all the parameters and data in a problem (Spiegelhalter et al., 1993), structured around conditional independence relationships motivated by the substance of that problem. In assessment, Bayes nets take the form of distributions for SM variables, for observable variables conditional on SM variables, and for parameters for population and conditional probability distributions. Psychometric models such as classical test theory (CTT), item response theory (IRT), factor analysis, and latent class models can be expressed as special cases of Bayes nets (Mislevy, 1994).

## The Relationship Between Claims and Student-Model Variables

Claims about students are expressed in substantively meaningful language. We turn to the language of mathematics for formal machinery to characterize and synthesize evidence that bears on claims. As we have said, the student model is the bridge between substantive claims and mathematical machinery for managing evidence and belief. The following paragraphs discuss four approaches for articulating a relationship between a set of claims and a probability distribution for one or more SM variables.

One approach is one-to-one relationship between a claim and a continuous SM variable. The claim concerns a student's overall proficiency with respect to some skill or knowledge, which is operationally defined as tendency to produce successful performances in a specified domain of tasks. There may be privileged points along this continuum, interpreted as "mastery" or "proficient" or "certifiable," in terms of which more specialized claims can be addressed, such as whether a student needs to review material or ought to receive a diploma. Behavioral objectives, from traditional guides on test design, are examples of claims often addressed using this approach. The SM variable is interpreted as a student's propensity to make performances at some level, perhaps as rated at several levels of quality. Observing performance in each task adds a nugget of evidence about the SM variable, the probability distribution of which constitutes grounding for a claim about the student's degree of proficiency.

A second approach comprises multiple claims and a single SM variable with a finite number of levels. Each value of the SM variable matches up one-to-one with a particular claim or set of claims. The ACTFL guidelines for language proficiency are an example. There are nine ordered levels in each of four separate scales, for Reading, Writing, Speaking, and Listening. The guidelines start at Novice Low, where a student can typically only use language in the modality in question in a few rudimentary ways, and go up to Advanced, where a student can use language fluently and in sophisticated ways. Each level is described by several statements about the kinds of things a typical student at that level can do, in situations with certain key features. Each such statement is a claim in its own right. The intention is that the statements within the description of a given level will typically go together well enough that a student can be characterized in terms of a single level, although there will be some performances above or below the most appropriate level. Note

that by partitioning language use into the four modalities, ACTFL does not directly address claims that would involve using modalities jointly, such as reading material then speaking about its meaning.

A third alternative for addressing multiple claims with a single SM variable is to model response probabilities in settings with particular key features. An example is the Document Literacy scale from the Young Adult Literacy Survey (YALS; Kirsch & Jungeblut, 1986), in which (a) open-ended tasks requiring simple written responses are generated according to features that are salient under the YALS cognitive model for processing documents; (b) a single IRT variable $\theta$ accounts for performance across all the tasks; and (c) a regression model using task features as predictors accounts for most of the variation in task difficulties. One can generate a whole family of (admittedly rather technical) claims using these features; for example: "Eighty-percent of the time, the student can do *feature matching* in documents that are *structured according to those features*, when *three* features need to be matched and there are *no close distractors* in the text." The italicized phrases are values for slots that can be filled in with values of task-model variables to generate a large number of claims. Because of properties (b) and (c), knowledge about a student's $\theta$ in terms of a current distribution can be mapped to degree of support for any of the claims.

A fourth approach directly tackles the interactions among competences and contexts: multiple SM variables can be called upon to express evidence for a claim. This is appropriate when multiple aspects of knowledge or skill are required in combination to support a claim, and students exhibit different patterns of proficiency in those skills. Separate SM variables are used to manage belief about these skills, which are conceptually distinguishable even though they may be called upon jointly to solve problems (e.g., knowledge of troubleshooting strategies and familiarity with a malfunctioning subsystem). Students' proficiency in a domain can be described in terms of constellations of competences they possess, and tasks can be described in terms of which competences they require—a multivariate generalization of the third approach above. A claim can thus be cast in terms of a student's proficiency with respect to a class of tasks with a given combination of salient features. The evidence for such a claim is expressed as the joint distribution for the SM variables corresponding to the skills these tasks call upon. In language assessment, this approach could be used to accumulate evidence for cross-modality

claims such as “Shu-Jing can express the gist of everyday conversations she hears in fluent speech, but not in writing at even the level of words or phrases.”

This fourth, multivariate, approach is not common in current assessment practice, but it is appealing in spheres of activity that call upon multiple aspects of knowledge or competence, such as language use. What is easy and what is difficult varies from student to student because tasks challenge aspects of competence in different combinations and to different degrees, in accordance with their features and requirements. If the claims we want to make require sorting out the reasons that people do well or poorly in settings with different kinds of features, we need a student model that is capable of making the necessary distinctions. This approach requires us to identify features of settings that stress different aspects of competence, and know how to sort out the evidence about the different aspects of competence in these complex situations. We will say more about these issues in the following sections on evidence and task models, and again in the section on assembly models when we discuss how the assessment designer manages the interplay among student, evidence, and task models to fill out an assessment argument.

**Example (Cont.): Student Models for the Language Assessment Examples**

Assessment arguments designed for Purpose 1 (placing out of the general academic writing class) and for Purpose 2 (satisfying an academic writing requirement for psychology) both concern a single proficiency aimed at a yes / no decision. In both cases, the language-use proficiency at issue is complex and multivariate, requiring high levels of both reading proficiency and writing proficiency. Purpose 2 further requires topical knowledge about psychology. But despite the fact that the proficiency is multivariate in both cases, a single student-model variable is sufficient machinery to serve the assessment purpose. Its values reflect the likelihood of high-quality performances in whatever situations correspond to the relevant claim.

We will use an IRT model with a single proficiency variable $\theta$ for Purposes 1 and 2. The basic idea is that probabilities for the observable variables from all tasks are functions of a student’s $\theta$ and one or more parameters characterizing each task. Figure 5 shows two representations of this kind of student model. The left panel shows a single variable (the oval labeled Overall Proficiency), with an associated

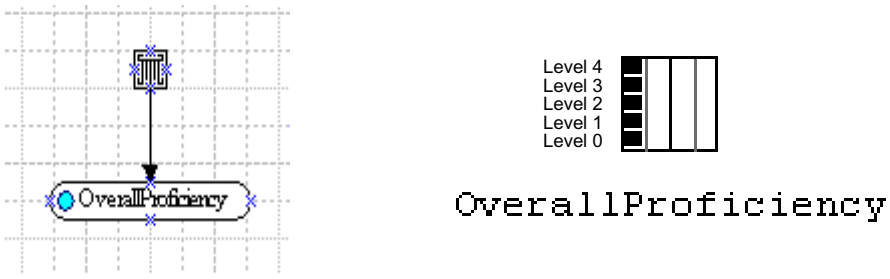



*Figure 5.* A single-variable student model.

probability distribution (the small matrix) that expresses current belief about the relative likelihood of its possible values.[8] The right panel shows the five discrete values which, for simplicity, we are using to approximate the continuous range usually used in IRT. This figure depicts a probability distribution starting with an uninformative distribution: probabilities of 1/5 for each possible value. These prior probabilities appear as proportionally scaled bars for the possible values. An informative prior could be used for a student about whom we had more information concerning the proficiency at issue, such as from previous performances or educational background. From whatever prior probabilities, evidence from a student's performances will be used to update the distribution.

Although the same statistical machinery is used for Purposes 1 and 2, the substantive interpretations of $\theta$ will differ, in accordance with the features of the tasks in the assessment, the kinds of performances the students produce, and the aspects of the performances that are evaluated. By controlling the demand for topical knowledge in the tasks constituting the two assessments—low and generic for Purpose 1, high and specific to psychology for Purpose 2—we induce different substantive interpretations to ground claims that suit their respective purposes.

Purpose 4 doesn't need a formal student model. Used by the student one task at a time, the practice assessment reports a rating for each individual essay, but any accumulation of evidence over essays is done outside the system, perhaps informally by the student or student's coach.

Purpose 3 addresses the same TLU as Purpose 1, but it requires a finer grained student model to support remediation. For students who cannot place out of

[8] This picture, and others in the same style, are screen shots from Educational Testing Service's Portal assessment design software.

academic writing, in which areas are they having trouble? We need to be able to say, for example, that Patricia can "read to learn" from the kinds of texts students encounter, but has difficulties integrating information across texts. We must have a student model with variables and values that can correspond to such statements. Figure 6 shows such a model. The left panel shows two SM variables, Reading and Writing. The left panel presents the view with both variables and distributions.[9] The right panel of Figure 6 shows the possible values of each variable, which we have put in correspondence with the Reading and Writing claims from Table 1. Again we depict uninformative prior distributions.

**Task Models**

A task model (TM) is a design object that bridges substantive considerations about the features of tasks that are necessary to elicit evidence about targeted aspects of proficiency, on the one hand, and on the other, the operational activities of authoring, calibrating, presenting, and coordinating particular assessment tasks.

A universe of individual tasks can be generated to accord with a given task model. The tasks differ on the surface, but they are variations on a thought-through argument for providing students the opportunity to demonstrate targeted proficiencies, in ways we will know how to interpret (e.g., Mislevy, Steinberg, &

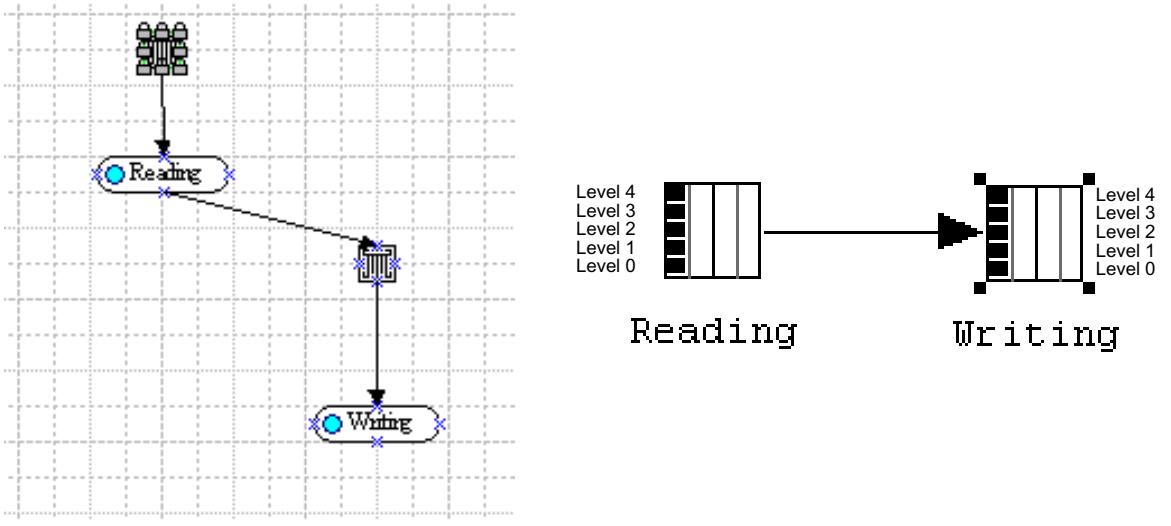



*Figure 6:* A two-variable student model.

[9] We expect Reading and Writing to be correlated in a population of students, so we have modeled Writing as dependent on Reading. There is no reason to prefer this direction here, but in other applications substantively indicated relationships such as prerequisition and inclusion suggest directions for modeling relationships among SM variables.

Almond, in press; Mislevy, Steinberg, Breyer, et al., in press). A task model contains two kinds of information. First, it specifies work products, or the form in which what the student says, does, or produces will be captured. Second, a task model contains a set of task-model variables that characterize key features of the environment in which the student will say, do, or produce something—characteristics of stimulus material, instructions, help, tools, and so on. Each task-model variable has a defined range of possible values. Each plays one or more roles, such as contributing to difficulty, focusing evidence on certain kinds of knowledge, facilitating task authoring, or guiding test assembly (Mislevy, Steinberg, & Almond, in press). In creating a particular task, a task author effects a value for each variable in the applicable task model, by providing materials or setting task conditions accordingly.

For a given assessment, one develops a set of task-model variables by simultaneously considering the substantive and the operational two points of view. With regard to substance, task models embody beliefs about the nature and structure of task situations, with regard to the kinds of performances, and in turn to the kinds of knowledge or proficiencies, about which they may be used to provide evidence (Mislevy, Steinberg, & Almond, in press).[10] Task-model variables formalize the notion of features of performance situations. To say "This task has Feature *x*, which corresponds to its having value *x* for TM variable *X*," is to say not only that it has feature *x* but that there are alternative features it might have had but does not—which correspond to other values of TM variable *X*. Further, feature *x* is present because it plays roles in the evidentiary argument for focusing evidence, affecting difficulty, heading off an alternative explanation of performance, and so on, as explicated in the task paradigm. For a particular task, then, the values of task-model variables constitute data for the evidentiary argument, characterizing the situation in which the student is saying, doing, or making something.

With regard to operations, some of the specifications and values of TM variables will guide the construction of tasks; others (perhaps an overlapping subset) will be used as information for calibrating, delivering, and selecting tasks in an operational assessment system. The interests of different actors, often specialists

[10] Statistical indices will play a role in flagging tasks that are unexpectedly hard or easy, possibly for reasons related to features of tasks we thought were incidental but now realize we must avoid, model, or rethink back at the level of substantive arguments. In iterative cycles of assessment design and tryouts, we expect to revise our substantive hypotheses, and accordingly our task models, in light of empirical data.

in different fields with responsibilities in different facets of assessment, come together at this point in assessment design. Constructing task models that integrate these disparate perspectives helps ensure coherence between substantive arguments and operational processes, and coordination among task construction, administration, and statistical analysis.

In the assessment design stage, then, task models are created and TM variables with their ranges of potential values are defined. Their connections with other elements of the assessment, in light of the overarching assessment argument, are worked through. In the task-authoring stage, individual tasks are created in accordance with a given task model, and their values on the relevant TM variables are determined. One way this happens is designating targeted values or ranges of values for certain TM variables, then creating tasks with features that are so encoded (e.g., Embretson, 1998). Another way involves designating the values of one set of TM variables a priori, but determining the values of the others from stimulus materials found or created later which satisfy the relevant values of the first set. Vocabulary and syntax codings of reading passages may be targeted roughly, for example, then coded exactly when specific suitable texts are identified. Another way to determine the values of TM variables for individual tasks is through negotiation. If literary analysis is the domain, and we need to observe a student's analysis of a literary work that the student is familiar with, it may be reasonable for the student to select a work—as long as it meets targeted values of other TM variables such as length, genre, or reading level.

Although indefinitely many tasks can be created in accordance with a given task model, the collection is constrained by the finite set of variables the task model contains, each with its defined range of possible values. The effects of other features of tasks, such as features of stimulus materials beyond those expressed as values of TM variables, become sources of uncertainty in the measurement models. When we fix values of selected task-model variables—depending on which variables we fix, the roles they play, and the value selected for any given variable—we can generate multiple families of more closely related tasks from a single task model. Different levels of generality or specificity can be chosen for task models, as suits the needs of the assessment project. There is no right level of detail, and one can envisage hierarchies of increasing specificity that are convenient for different participants in an assessment system to work with. This can be a highly structured process for a large-scale assessment, in which teams are charged with creating and filling in task

models to a point at which other teams write individual tasks, pre-calibrate IRT models, and assemble presentation materials. The structuring could be built into an assessment-authoring tool, where groups of master teachers design templates, or partially filled-in task models, for classroom teachers to customize. The same kind of thinking happens implicitly when a student plans a project with her teacher, to make sure it has features that will provide the kind of evidence that need to be manifest.

**Example (Cont.): A Task Model for the Language Assessment Examples**

We sketch a single, broadly defined, task model that could be used to generate tasks for all four of the language assessment purposes in our example. It concerns a student reading one or more academic texts with certain features, then producing a response in written English that meets specified requirements. A partial set of salient features, motivated by analyses of the TLUs and a fore-knowledge of the purposes of the assessments, gives rise to a collection of TM variables that can be used to generate tasks that serve those different purposes. These TM variables are important because they can be used to elicit evidence about particular aspects of knowledge or proficiency. We will not discuss additional TM variables one might want to define to further characterize domains of tasks that could be generated, nor will we detail specifications for work products, administration conditions, and other requirements that would be needed for a fully specified task model. The specifications would include the form in which the student produces text and how it will be recorded and stored, as well as tools and support that will be made available (e.g., dictionaries, reference materials, spell checkers).

The TM variables we do address are central to the conceptual argument about academic reading and writing proficiency. They concern the number and nature of texts, and the nature and expectations of the writing requirements. The variables and their values are motivated by the features of situations that are necessary to get evidence to back claims about students' academic reading and writing proficiencies, as shaped by the four assessment purposes. Table 2 lists some key TM variables, which are accompanied by minimal sets of possible values and comments about their relationships to potential student-model and observable variables. In most cases the values are specified as "low, medium, high." In an implemented assessment, these values would need to be defined operationally or algorithmically,

Table 2

Sketches of Illustrative Task-Model Variables

| Variable | Values | Comments |
|---|---|---|
| Number of rhetorical frames in stimuli | 1, >1 | Must be >1 to obtain evidence of reading to integrate. |
| Status of rhetorical frames for stimuli | None, explicit, implicit | "Stimulus frames" does not apply to sentence-level texts. A text with explicit frames is needed to get evidence about reading for basic information or reading to learn. Implicit frames that span texts are needed to get evidence of reading to integrate. Could define finer grained TM variables to address which frames are used, which would be needed to support claims about proficiency with specific frames. |
| Stimulus length | Sentence, paragraph, multi-paragraph, extended | Affects difficulty, with respect to cognitive load. |
| Stimulus syntax | Simple, moderate, complex | Affects difficulty, with respect to grammatical competence and cognitive load. |
| Stimulus vocabulary | Simple, moderate, complex | Affects difficulty, with respect to vocabulary competence and cognitive load. This is distinct from topical demand of vocabulary. |
| Topical knowledge demand | Low, medium, high | Values should be medium or high when topical knowledge is at issue. Should be low when topical knowledge is not at issue, so it won't be an alternative explanation of poor performance. Could define finer grained TM variables to address kinds or areas of topical knowledge; would need to if claims include topical knowledge. |
| Response frame | Not applicable; recognize given frame; manipulate given frame, generate frame | What is the student informed, either immediately or through previous instruction and examples, about use of an organizing frame in a response? Could be irrelevant, or recognizing or working with frames that are provided, or student must generate one. Must be at least *recognize* to get evidence about reading to learn; must be *generate* to get evidence about reading to integrate. |
| Response length | Discrete, sentence level, paragraph level, extended | Note that values correspond to kinds of texts that must be observed in order to obtain evidence about students' use of conventions and techniques for academic writing at various levels. |
| Response register | Not applicable, formal, informal | What is the student informed about expectations with respect to register in the written response? This is not relevant for discrete responses. Use of academic register is an aspect of sociolinguistic competence. |
| Response syntax | Not applicable, relevant | Is the student informed that syntax will be evaluated? In an assessment that focuses on topical knowledge, it may not be. |

or illustrated with enough examples that people would agree on their application to specific tasks.

Working with the ranges of values defined by these TM variables, we now describe families of tasks that share a similar structure (i.e., conform with, or are generated from, the same task model), but because they vary the values of key TM variables, provide evidence about different combinations of proficiencies at different levels. For Purpose 1, we want tasks that meet the qualifications shown in the second column of Table 3. We will call this Task Family 1. These settings are motivated by considerations from Domain Modeling, concerning task features that are needed to elicit evidence about the general academic reading/writing proficiency that Purpose 1 targets. The ranges of TM variables specified in the Task Family 1 describe tasks that can evidence the highest levels of the finer grained academic reading and writing SM variables, but not topical knowledge. For Purpose 4, which is providing students with approximate ratings on practice essays before taking the Purpose 1 assessment, exactly the same values of TM variables and the same resulting tasks are appropriate, even though operational processes such as evaluation, measurement, and delivery will be quite different.

For Purpose 2 we want tasks that have the same ranges of TM variables except that Topical Knowledge Demand must be High, specifically in the domain of psychology. Call this Task Family 2. In both Task Family 1 and Task Family 2, the selected subset of values of TM variables to be used will focus on evidentiary potential for a combined high level of proficiency across all of the aspects of proficiency that are involved.

Purpose 3 required a more variegated set of claims in order to support instructional recommendations. We can conceive of different combinations of task features that focus attention at different levels of proficiency in academic reading and writing, so that we will get some information directly about all of the combinations of proficiencies students might have in order to ground claims suggesting different courses of instruction. We don't want to stress topical knowledge about psychology or anything else, so we require Topical Knowledge Demand to be Low.

Here are three specific combinations we will use again as illustrations in the sections on measurement models and delivery systems. The designated values or ranges of TM variables are shown in the appropriate columns of Table 3.

Table 3

Task-Model Variable Assignments for Illustrative Task Models

| | Required value or range | | | | |
|---|---|---|---|---|---|
| Variable | Task Family 1 | Task Family 2 | Task Family 3a | Task Family 3b | Task Family 3c |
| Number of rhetorical frames in stimuli | >1 | >1 | 1 | >1 | >1 |
| Status of rhetorical frames for stimuli | Implicit | Implicit | None or explicit | Implicit | Explicit |
| Stimulus length | Paragraph, multi-paragraph, or extended | Paragraph, multi-paragraph, or extended | Sentence or paragraph | Paragraph, multi-paragraph, or extended | Paragraph |
| Stimulus syntax | Moderate or complex | Moderate or complex | Simple or moderate | Moderate or complex | Moderate |
| Stimulus vocabulary | Moderate or complex | Moderate or complex | Simple or moderate | Moderate or complex | Moderate |
| Topical knowledge demand | Low | High | Low | Low | Low |
| Response frame | Generate frame | Generate frame | Not applicable | Not applicable | Generate frame |
| Response length | Extended | Extended | Discrete | Discrete | Extended |
| Response register | Formal | Formal | Not applicable | Not applicable | Formal |
| Response syntax | Relevant | Relevant | Not applicable | Not applicable | Relevant |

- Task Family 3a. Tasks in this family have both simple reading demands and simple writing demands. Performances on these tasks have the potential to provide evidence about whether students can perform at the most basic levels of academic reading and writing, but not much above that.

- Task Family 3b. Tasks in this family have advanced reading demands but simple writing demands. That is, challenging stimulus materials are presented, and the student will need to generate a synthesizing frame, but only discrete and fragmented writing is required in doing so. Performances on these tasks can tell us about whether students can perform at high levels of academic reading, but not much about their productive academic writing proficiencies.

- Task Family 3c. Tasks in this family have moderate reading demands but advanced writing demands. Performances on these tasks can tell us whether students can perform at the moderate levels of academic reading and advanced levels of academic writing.

We have used TM variables to manage the kinds of observations it will be possible to make, and thus the kinds of knowledge and proficiencies we will be able to get evidence about. This is building into tasks the potential to obtain relevant evidence from work products. We have yet to discuss the form in which we will extract nuggets of evidence (i.e., observable variables), how we will synthesize evidence in terms of students' knowledge or accomplishments more broadly (i.e., student-model variables), or how we will connect the two. These are issues that are the focus of evidence models.

**Evidence Models**

An evidence model lays out the part of the evidentiary argument that concerns reasoning from the observations in a given task situation to revised beliefs about student-model variables. There are two components in an evidence model. The evaluation component contains rules for extracting nuggets of evidence from individual performances, as values of observable variables. The measurement component contains statistical models for synthesizing the information from observable variables across performances in terms of belief about values of SM variables. Stated more technically, the evaluation component specifies procedures for ascertaining the values of observable variables from students' work products, and the measurement component specifies models for constructing likelihood functions for SM variables, which are induced by the values of the observed variables.

Structuring evidence models is a step of formalization and specification. We are moving from evidence paradigms, which are fuller in terms of substance but sketchier in terms of operational elements, to specifications for pieces of machinery: evaluation algorithms or rubrics in the evaluation component, and structures and parameterizations for the statistical models in the measurement component.

**The Evaluation Component**

The evaluation component of an evidence model might be called task-level scoring. The *nature* of the desired observations was determined in domain modeling—an indication that the student could recognize, generate, or fill in an

organizing frame for comparing two reading texts, for instance. Particular forms, evaluation procedures, and formal definitions of observable variables are now specified.

The evaluation component, then, indicates how one identifies and evaluates the salient features of a performance, and expresses them as values of observable variables. A work product is a unique human production, be it as simple as a response to a multiple-choice item or as complex as repeated cycles of treating and evaluating patients in a medical simulation. Observable variables provide, in a common structure for all students, evaluative summaries of the features the assessment designer has deemed important for the assessment's purpose. Evaluation rules examine what a student says, does, or makes in terms of these observables, and sets each one to a value reflecting a particular evaluative outcome. They can be automated, demand human judgment, or require both in combination. In some assessments the specifics of a student's work and its evaluation are negotiated between examinees and examiners, within a broader framework of valuation. Doctoral dissertations and the Concentration section of the AP Studio Art Portfolio assessment are examples.

Depending on assessment purpose, different aspects of a given performance can be captured as work products. Further, as we will see in the language assessment example, different aspects of the same work product can be evaluated as observable variables. A short impromptu speech, for example, potentially contains information about a student's subject matter knowledge, presentation capabilities, and English language proficiency. Any of these and more, or any combination of them, could be the basis of observable variables. Which are pertinent, what aspects of proficiency they reflect, and why they are relevant to the assessment argument will have been worked through in Domain Modeling.

As part of the backing for the warrant from which we reason from students' performances to their knowledge or proficiencies, assessors must generally inform students of the aspects of a performance that are being evaluated, and by what criteria. That a student did not understand how their work would be scored is an alternative hypothesis for poor performance that we can and should avoid. Worked-through examples and in-depth discussions of rubrics, for example, are as important to students and teachers in classrooms as they are to raters in large-scale scoring systems. Situative psychology emphasizes the importance of students coming to

understand, even helping to create, evaluation procedures in developing competence (Greeno, Collins, & Resnick, 1997).

**Example (Cont.): Evaluation Rules for the Language Assessment Examples**

The four assessment purposes in our running language assessment examples are all based on the same TLUs. In all four cases, we can generate tasks from the same task model using the task families we obtained by constraining the values of certain TM variables to suit the differing purposes. In all cases, the work products are texts students write, simple or extended, as required. In view of the differing purposes, however, we can choose to define observable variables of several kinds.

The first two observable variables we consider are holistic ratings of performances. Holistic ratings are a common choice for assessments aimed at purposes like our Purposes 1 and 2, which are meant to support decisions based on overall levels of performance on tasks in a given domain. They address the degree to which a student succeeds in accomplishing the goal specified in the task. For Purpose 1, a successful performance is grammatically sound and stylistically appropriate; it provides a suitable unifying frame for the information in the texts; and correctly integrates the important information from those texts into that frame. A 4-point rating scale that could be used for this purpose, labeled *effectiveness of response,* is shown in Table 4.[11] A modification of this scale for Purpose 2 would

Table 4

Rubric for Evaluating an Overall Effectiveness of Response

| Rating | Name | Description |
|---|---|---|
| 0 | None | Little or no indication of understanding of texts or expression of information in them. |
| 1 | Limited | Some degree of cross-text relationships, although incomplete; some structure provided, but with serious shortcomings as to grammar or structure. |
| 2 | Moderate | Obvious structuring of appropriate information, although with lapses and errors in expression or understanding. |
| 3 | Complete | Grammatically sound and stylistically appropriate essay, providing a suitable unifying frame and correctly integrating the important information from the texts. |

[11] Rating scales like this one and the others that follow are incomplete evaluation rules without rated examples, training procedures, or both. Their meaning is constructed—by raters, instructors, and students alike—through a process of socialization.

additionally include the effective application of topical knowledge, in this case information or organizing schemes from the domain of psychology that are not present in the stimulus materials.

Bachman and Palmer (1996, p. 275 ff) provide five more focused rating scales with similar definitions that could be applied to integrative essays, addressing *syntax, vocabulary, rhetorical organization, cohesion, and register*.[12] *Syntax* and *vocabulary* can be used with responses that require only discrete or sentence-level responses. *Rhetorical organization, cohesion,* and *register* can only be evidenced in performances above the sentence level, for they concern the use of conventions and devices for organizing extended discourse. In other examples they provide a similar rating scale for Use of Topical Knowledge. To give a student an opportunity to produce performances that can be evaluated with respect to Use of Topical Knowledge, task features that must be present are (a) stimulus materials that present information related to domain-relevant relationships that are not included in those materials, and (b) instructions that make it clear to the examinee that producing such relationships is required. The first of these would be expressed as a value of a task-model variable, as discussed in the section on task models. The second, an aspect of the instructions provided to the student, would have been specified in the conditions of administration, also part of the task model.

The point to be made in this example is not merely that performance can be evaluated for different features. It is that one considers beforehand which performance features are needed as evidence for targeted proficiencies, and what task features must be present in order to elicit that evidence. The necessary task features are expressed as values of TM variables, present in the TM because their importance in the evidentiary argument was established in Domain Modeling.

Rating scales, counts, or flags could be defined for far more detailed and focused aspects of performances. So too would be a number of grammatical and word use features that would be required for the automated reading that appears in the assessment for Purpose 4. Fine-grained observable variables such as these might be necessary for a purpose such as computer-based instruction as well. The reader

---

[12] Bachman and Palmer give them names like *Knowledge of Syntax* and *Knowledge of Rhetorical Organization*. Although it is clear that these rubrics describe qualities of specific performances, some ambiguity is introduced by using names that sound like qualities of students. Bachman and Palmer are looking ahead toward one-to-one relationships between these rating scales as descriptors of performances and descriptors of students' propensities to produce performances at various levels. This style of naming observables is less apt for the many-to-many relationships discussed below.

interested in these fine-grained characteristics of written texts or in automated scoring of essays is referred to Frase et al. (in press).

**The Measurement Component**

The *measurement component* of the evidence model contains statistical models for how the observable variables depend, in probabilistic terms, on student-model variables. A measurement model affords reverse reasoning through the machinery of probability: in the realm of machinery, from observable variables to student-model variables; in the realm of substance, from observations to claims. In a nutshell, the rationale is this:

Student-model variables are posited to "explain" performance, in accordance with the purposes and substance of the assessment. Knowledge about the values of the SM variables at any point in time is represented by a probability distribution over their values. Conditional probability distributions, from SM variables to the observables of each task, are an expression of a warrant in its forward direction. These conditional probabilities are procured, from some combination of experience and theory; that is, pretesting, expert judgment, and/or model-based parameterization in terms of TM variables. Responses to different tasks are modeled as conditionally independent from those of other tasks, given the requisite SM variables. (Observations within tasks may be modeled as conditionally dependent, however, as we will see in our language testing examples.) When observations are obtained for any task, they induce a likelihood function for SM variables, which may be used to revise belief about SM variables using Bayes Theorem to produce a posterior probability distribution for the SM variables.

In order to describe more formally how Bayesian updating takes place and to set the stage for more elaborated measurement models, we will look more closely at Rasch's (1960/1980) IRT model for right/wrong test items. Under the Rasch model, the probability of a correct response takes the form

$$\text{Prob}(X_{ij}=1|\theta_i,\beta_j) = \Psi(\theta_i - \beta_j),$$

where $X_{ij}$ is the response of Student $i$ to Item $j$, 1 if right and 0 if wrong; $\theta_i$ is the proficiency parameter of Student $i$; $\beta_j$ is the difficulty parameter of Item $j$; and $\Psi(\cdot)$ is the logistic function, $\Psi(x) = \exp(x)/[1+\exp(x)]$. The probability of an incorrect response is $1-\Psi(\theta_i - \beta_j)$. In words, the probability of a correct response is near zero for a student with a very low value of $\theta$; it increases gradually, up to 50-50 chances for

students with $\theta$ at the same value as the item's difficulty parameter; and for $\theta$'s increasingly higher, the probability of a correct response approaches one.

Letting $p(x_j|\theta)$ denote the conditional probability of a particular response value $x_j$, we use the assumption[13] of conditional independence to write the probability of the responses $(x_1, x_2, ..., x_n)$ to a collection of *n* items as

$$p(x_1, x_2, ..., x_n|\theta) = \prod_j p(x_j|\theta)$$

The joint distribution of $\theta$ and $(x_1, x_2, ..., x_n)$ is then

$$p(x_1, x_2, ..., x_n, \theta) = p(\theta)\prod_j p(x_j|\theta)$$

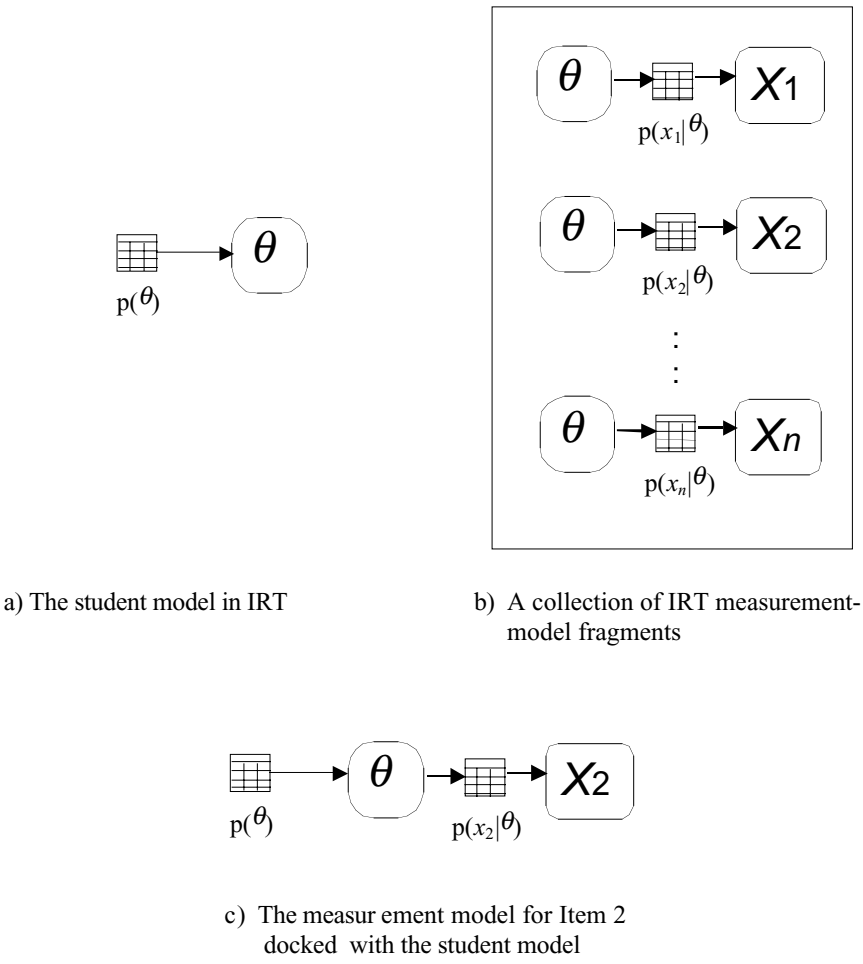
Once particular response values are ascertained, say $(x_1^*, x_2^*, ..., x_n^*)$, the posterior distribution for $\theta_i$ is obtained via Bayes theorem as

$$p(\theta|x_1^*, x_2^*, ..., x_n^*) \propto p(\theta)\prod_j p(x_j^*|\theta) = \frac{p(\theta)\prod_j p(x_j^*|\theta)}{\int p(\phi)\prod_j p(x_j^*|\phi)d\phi}.$$

Figure 7 depicts the IRT statistical model graphically. Panel "a" is the student model, which is just $\theta$ and its associated probability distribution. Panel "b" is a collection of IRT measurement-model fragments, one per item. Each fragment represents $p(x_j|\theta)$, a conditional distribution that is determined by the item parameter of Item *j*. Panel "c" shows the measurement-model fragment for an item "docked" with the student model to produce a joint probability distribution for $\theta$ and the response to Item *j*. From here one can carry out reverse reasoning, via Bayes theorem one item at a time, to update belief about $\theta$ when the value of the response becomes known. This structure is the basis of adaptive testing; items are selected one at a time to provide most information about $\theta$, given responses that have been observed so far.

---

[13] Again, conditional independence per se is not a belief about psychology or purposes of assessment, but a conjecture about relationships among the variables we've defined to organize our thinking. There are several strategies to achieve correspondence between the patterns our model can express and the patterns we perceive among real-world behaviors. We use what we know about theory and experience to suggest the form of models, be they simple like IRT or more complex. We arrange observational settings to be compatible with this structure, designing tasks to focus on targeted knowledge and minimize extraneous factors. And we use model-fit analyses to flag instances when data differs meaningfully from the posited structures.




a) The student model in IRT

b) A collection of IRT measurement-model fragments

c) The measur ement model for Item 2 docked with the student model

*Figure 7.* A graphical depiction of the statistical model for IRT adaptive testing.

IRT adaptive testing is a straightforward application of the ideas from probability-based reasoning that open the door to more flexible, yet rigorous, modeling in assessment. Factor analysis, latent class models, generalizability theory, and multidimensional scaling are also special cases of the same kind of reasoning. Each addresses particular kinds of data, SM variables, and relationships between them. We can extend these ideas and these structures to suit the nature of the student model and observable variables in any given application (Almond & Mislevy, 1999). We noted previously that $\theta$ can be vector-valued. The observable variable obtained by evaluating performance on Task $j$, say $X_j$, can be vector-valued as well.

We generally construct tasks so that there will be conditional independence of responses to different tasks. In complex situations, different subsets of multivariate SM variables are posited to influence different responses, and observable variables

that reflect multiple aspects of the same complex response may be modeled as dependent beyond the associations caused by their SM parents (i.e., the SM variables they are modeled as dependent on). We are fortunate in assessment that to an extent greater than in fields such as history or jurisprudence, we can shape the situations in which we will gather data. Through our design work on task paradigms, then detailed in task models, we can highlight the patterns that are important to our purposes (which we build into our measurement models) and minimize those that are irrelevant (which constitute unmodeled sources of variation, or noise, with respect to the measurement models).

The statistical model used in the measurement component of the evidence model may extend to aspects of the evaluation component, as when judgments are required to ascertain the values of observable variables from complex performances. Issues of rater accuracy, agreement, leniency, and optimal rating designs can be addressed with a measurement model that includes task evaluation as well as *evidence accumulation* (see Brennan, 1983, and Cronbach, Gleser, Nanda, & Rajaratnam, 1972, for a generalizability-theory perspective on these issues, and Linacre, 1989, and Patz & Junker, 1999, for an IRT perspective).

### Bridging Student Models and Task Models With Structured Measurement Models

Two active areas in psychometrics are extending IRT to multivariate student models and incorporating task-feature/student-proficiency relationships formally into measurement models (Pellegrino, Chudowsky, & Glaser, 2001). With multivariate generalizations of IRT, one can extend to situations in which multiple aspects of knowledge and skill are required in different mixes in different tasks. One stream of research on multivariate IRT follows the tradition of factor analysis, using analogous models and focusing on estimating structures from tests more or less as the items come from the test developers, by whatever processes, for whatever purposes, happen to have been in place (e.g., Reckase, 1985). A contrasting stream starts from multivariate conceptions of knowledge and constructs tasks that contain evidence of that knowledge in theory-driven ways (e.g., Adams, Wilson, & Wang, 1997). This structural extension of IRT fits neatly with a structural perspective on task construction.

Embretson (1983) argued for incorporating task-model variables into the statistical model, thus making explicit the ways that features of tasks impact

examinees' performance and constructing tasks around these features. This is important because it forges stronger connections among the substantive, statistical, and operational facets of assessment. In the ECD framework, the substantive argument connecting task features and statistical models is built into the design objects of the CAF; specifically, the conditional probabilities for observable variables given SM variables are modeled as depending in part upon the values of task-model variables. The tools of probability-based reasoning thus become available to examine how well substantively motivated beliefs about student/task relationships hold up in practice (validity), how they affect measurement precision (reliability), how they can be varied while maintaining a focus on targeted knowledge (comparability), and whether some items prove hard or easy for unintended reasons (fairness).

A signal development in this regard was the linear logistic test model, or LLTM (Fischer, 1973; Scheiblechner, 1972). The LLTM is an extension of the Rasch model shown above, with the further requirement that each item difficulty parameter is the sum of effects that depend on the features of that particular item:

$$\beta_j = \sum_k q_{jk} h_k$$

where $h_k$ is the contribution to item difficulty from Feature $k$, and $q_{jk}$ is the extent to which Feature $k$ is represented in Item $j$. In terms of the ECD structures, $\beta_j$ determines conditional probabilities in the measurement model for Item $j$, and $q_{jk}$ is a known quantity determined by the values of one or more task-model variables as they are realized in Item $j$. Embretson (1998) walks through a detailed example of test design, psychometric modeling, and construct validation from this point of view. Several authors describe multivariate student models that formally model the relationship between SM variables and observable variables in terms of task-model variables, further extending the LLTM (e.g., DiBello, Stout, & Roussos, 1995; Falmagne & Doignon, 1988; Pirolli & Wilson, 1998; and Tatsuoka, 1990).

These kinds of models are not widely used at present, in large part because they are unfamiliar and difficult to use—but also because there is little payoff for using them unless tasks are developed jointly with the measurement model, so the complex model can be used to support inferences in terms of preplanned and substantively important patterns in data that simpler models can't express. Advances in statistical computing are beginning to ease the difficult-to-use problem. These developments concern resampling-based estimation, full Bayesian analysis,

and modular construction of statistical models (Gelman, Carlin, Stern, & Rubin, 1995). These developments are softening the boundaries between researchers who study psychometric modeling and those who address the substantive aspects of assessment. Some of our own work along these lines appears in Mislevy et al. (in review); Mislevy, Almond, Yan, and Steinberg (1999); and Mislevy, Steinberg, Breyer, Almond, & Johnson (in press).

**Dealing With Proficiencies That Have Many Aspects**

Assessors have resorted to three approaches when students vary substantially along multiple dimensions, such as language proficiency. The first is choosing focused aspects of competence or performance, and managing task features so as to challenge only those aspects while holding demands on other aspects below thresholds that the students of interest almost surely surpass. This approach leads to tests that each focus on a single well-defined aspect of competence or performance. This is how variation with respect to non-language competences are handled in tests of grammar and vocabulary: They minimize the need for sociolinguistic and pragmatic competences. One can assess a wide array of competences separately using this approach and use a familiar unidimensional model such as CTT or IRT to gauge proficiency in each competence in turn. However, this approach misses much of the interplay among skills that takes place in actual use. It solves the low generalizability problem, though at the expense of authenticity.

The second approach is administering complex tasks, evaluating success in performance, and accumulating an overall proficiency by averaging over all the many task-specific sources of variation. This approach captures a general rating of success in a task and again uses a unidimensional model to measure a general proficiency in the domain of tasks. The variation from one examinee to another as to which of these features prove troublesome and which do not is averaged over, as measurement error. This solution provides better authenticity than discrete skills assessments, but risks the low generalizability problem when different tasks call for different mixes of competences and students have differing profiles in those respects.

The third approach, used by far the least often, is modeling the variation in students and tasks at some level with multivariate models. Each student is characterized by more than one variable, each reflecting a distinct aspect of proficiency, and each task is characterized by the degree to which it tends to stress

the different aspects of proficiency. Now student-by-task interactions that render different tasks easy for some students and hard for others can be modeled and expressed as differing profiles of proficiency among students, as opposed to lost as measurement error in a univariate model. This approach addresses both the generalizability problem and the authenticity problem, but at the cost of complexity and unfamiliarity.

Which approach should one take? It depends on the nature of the claims one wants to support. We will look next at examples from language assessment, where the choice can be seen as driven by the assessment purpose. To anticipate, the second approach is satisfactory for Purposes 1 and 2, but the third approach is best suited to Purpose 3.

### Example (Cont.): Measurement Models for the Language Assessment

In each of the tasks generated under our academic reading / writing task model, a student must read one or more pieces of stimulus material, then produce a written response in accordance with a directive. This section describes two measurement models, one for Purposes 1 and 2 and the other to Purpose 3. The conditional probability matrices appearing in these examples are hypothetical and simplified as much as possible to illustrate key points. In particular, we have not shown the parameterization of conditional probability matrices in terms of task features. This structure adds an additional layer in the statistical model while providing more parsimonious expressions for conditional probabilities, along the lines of the Fischer (1973) and Adams, Wilson, and Wang (1997) models mentioned earlier (for examples see Almond et al., 2001, and Mislevy et al., in review).

### A Measurement Model With One SM Variable and One Observable Per Task

The modeling approach we illustrate for Purposes 1 and 2 is by far the most common in practice: A single observable variable is extracted from each performance, and a single student-model variable is operationally defined as the tendency to perform at higher rather than lower levels. Its simplest form comprises right / wrong multiple-choice test items, number right as a student's score, and classical test theory as a measurement model. We'll look at a more elaborated version of the same one-to-one structure: A single holistic rating of each essay and graded-response IRT $\theta$s for students' overall proficiencies.

Consider an assessment aimed at Purpose 1, consisting of three tasks from Task Family 1. Each requires integrating information (which we *cannot* observe) across multiple pieces of stimulus material and exhibiting this integration in a formal written essay (in which we *can* observe the use of writing conventions and clues as to whether integration has occurred). A single observable is evaluated from each performance, the four-point overall *effectiveness of response.* The student model contains one SM variable, *academic writing proficiency*, which encompasses both academic reading and writing. It is operationally defined by the task features and holistic-rating observable variables across tasks. $\theta$ corresponds to a student's propensity to provide higher quality responses on these tasks, in this setting, evaluated in this way.

Figure 8 shows the evidence-model Bayes net fragment for a Family 1 task. It would be docked with the student-model fragment shown previously as Figure 5. Table 5 gives a hypothetical conditional probability matrix for one task, then works through the steps of an instance of Bayesian updating for the response to just this one task. The steps are interpreted as follows:

- Panel "a" gives a prior distribution for $\theta$, or the distribution that expresses our belief about a student's value of $\theta$ before we learn the rating for this task, $X$.

- Panel "b" gives the conditional distributions for $X$, corresponding to each possible value of $\theta$. These distributions, read across rows, are the probabilities for the responses from a student known to be at that level. For a student for whom $\theta$ is Level 4, the probability of a *complete response* is .73. The probabilities of *moderate, limited,* and *none* responses are .15, .07, and .05

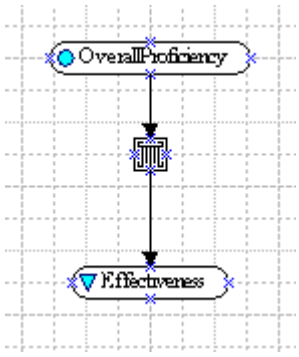



*Figure 8.* A one-to-one correspondence between student- and observable variables.

Table 5

Bayesian Calculations for "Overall Academic Writing Proficiency" and a Holistic Rating of Effectiveness for a Representative Task-Family 1 Task

a) Prior Distribution for $\theta$

| $\theta$ | Probability |
|---|---|
| Level 4 | 0.20 |
| Level 3 | 0.30 |
| Level 2 | 0.30 |
| Level 1 | 0.15 |
| Level 0 | 0.05 |

b) Conditional Probability of Response Rating

| $\theta$ | Complete | Moderate | Limited | None |
|---|---|---|---|---|
| Level 4 | 0.73 | 0.15 | 0.07 | 0.05 |
| Level 3 | 0.50 | 0.23 | 0.15 | 0.12 |
| Level 2 | 0.27 | 0.23 | 0.23 | 0.27 |
| Level 1 | 0.12 | 0.15 | 0.23 | 0.50 |
| Level 0 | 0.05 | 0.07 | 0.15 | 0.73 |

c) Likelihood Induced by Observed a “Complete” Response

| $\theta$ | Complete | Moderate | Limited | None |
|---|---|---|---|---|
| Level 4 | 0.73 | 0.15 | 0.07 | 0.05 |
| Level 3 | 0.50 | 0.23 | 0.15 | 0.12 |
| Level 2 | 0.27 | 0.23 | 0.23 | 0.27 |
| Level 1 | 0.12 | 0.15 | 0.23 | 0.50 |
| Level 0 | 0.05 | 0.07 | 0.15 | 0.73 |

d) Prior x Likelihood

| $\theta$ | Prior x Likelihood |
|---|---|
| Level 4 | 0.146 |
| Level 3 | 0.150 |
| Level 2 | 0.081 |
| Level 1 | 0.018 |
| Level 0 | 0.003 |

e) Posterior Distribution

| $\theta$ | Probability |
|---|---|
| Level 4 | 0.367 |
| Level 3 | 0.377 |
| Level 2 | 0.204 |
| Level 1 | 0.045 |
| Level 0 | 0.007 |

respectively. These conditional probabilities might have arisen from expert opinion, fitting a measurement model to pretest data, approximating probabilities based on features of the task, or a controlled study of students at known levels of proficiency. They reflect the forward reasoning that characterizes a warrant; *if* $\theta$ is at such-and-such a level, *then* the probabilities of making each rating in a response are such-and-such.

- Panel "c" contains the same numbers as Panel "b," but reflects the perspective after observing a response rating of *complete*. The highlighted column gives, in each row, the probability of a complete response at that level of $\theta$. This is the likelihood vector for $\theta$ induced by observing a complete response, the proportions by which we should revise our belief about the student's $\theta$. The likelihood vector reflects information for reasoning in the reverse direction of the warrant that the conditional probability matrix represents.

- Panel "d" is the first step in applying Bayes theorem: At each value of $\theta$, the prior probabilities from Panel "a" are multiplied by the likelihood values from Panel "c." The results are in the correct proportion to our revised beliefs about $\theta$, but they are not probabilities yet because they do not sum to one.

- Panel "e" completes Bayes theorem by dividing each of the values in Panel "d" by their sum to produce the posterior distribution for $\theta$. Comparing this posterior to the prior in Panel "a," we see that observing a Complete response has caused our beliefs to move toward higher levels of *overall writing proficiency*. This posterior can now be used as the prior for accumulating additional evidence, whether it is from the other two tasks from Task Family 1 or a different kind of task with different kinds of observables that also have Overall Writing Proficiency as a parent.

Over several tasks, some variation among a student's performances would be expected. Just how much variation is typical within response vectors, and how much the tasks differ from one another as to difficulty and amount of information, is embodied in the conditional probabilities and can be estimated from responses. The information from each observable induces a likelihood, and accumulating them in the manner of Table 5 synthesizes their combined effect for belief about their SM progenitors. The bars in Figure 9 depict belief about $\theta$ resulting from observing three responses, evaluated as Moderate, Complete, and Moderate, respectively, yielding a posterior that is concentrated around Level 3.

The same statistical machinery could be used for an assessment aimed at Purpose 2, using tasks generated under Task Family 2. The nature of the observational setting has changed, however, so that Topical Knowledge Demand is fixed at High; the evaluation rules have changed, so that a high-quality response must additionally exhibit an advanced understanding of psychology; and the interpretation of $\theta$ is now a proficiency that combines knowledge of psychology as well as being able to reason integratively and write using academic writing conventions with respect to psychology content.

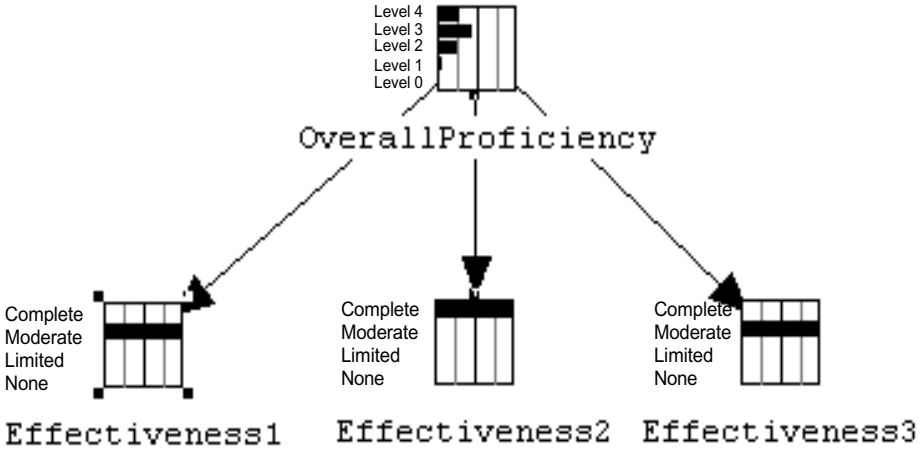



*Figure 9.* Updated probabilities under the one-to-one model.

## A Measurement Model With Multiple SM Variables and Multiple Observable Per Task

The performances elicited by the tasks we have designed for Purpose 3 tap reading and writing proficiency jointly, but with varying amounts of stress on each depending on the features built into each task. This section describes a measurement model with two SM variables, for reading and writing, and multiple observables per task, for different aspects of each task performance. The conditional probability distributions are structured around what we know about task features and their relationship to the two proficiencies in such a way as to disambiguate evidence about the SM variables contained in the multiple and overlapping observable variables.[14]

Figure 10 depicts such a model, as applied with one task from Task Family 1. The two SM variables on the left, *reading* and *writing,* are shown as parents of the observable variables on the right. There is an additional parent of all of the observables for this task (and no other), labeled *TaskContextj,* introduced to handle dependencies among multiple rated aspects of the same task, which we will discuss shortly. Some of the observable variables, such as Syntax, depend on Writing but not Reading. Whether a student understood the information in a stimulus text doesn't bear directly on whether the sentences the student produces are grammatical.

[14] Adams, Wilson, and Wang (1997), call these 'within-items" models.

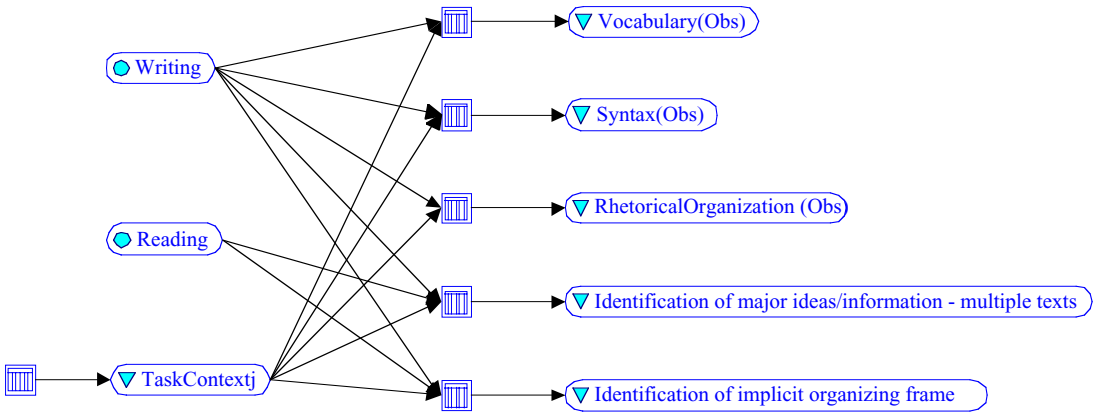



*Figure 10.* A many-to-many correspondence between student-model and observable variables for one task.

Vocabulary, Syntax, and Rhetorical Organization also have this property. On the other hand, the observable variables reflecting comprehension, namely Identification of Information and Identification of Implicit Frame, have both Reading and Writing as parents. Writing is necessary in this task to produce evidence about what a student has understood. Failure to produce a comprehensible response thus admits to multiple explanations: Reading proficiency is low; writing proficiency is below the hurdle necessary to communicate what one has comprehended; or both. Reverse reasoning through a probability matrix incorporating this structure affects the nature and strength for revising beliefs about a student's reading and writing proficiencies after seeing such a response.

Returning to $\text{TaskContext}_j$ and conditional dependence: Multiple-rated aspects of a complex performance are subject to task-specific effects such as distractions or time limit pressures, familiarity with context or content, or misunderstanding of expectations or stimulus materials. Failing to deal with this phenomenon results in overstating the evidence about SM variables; the effect is analogous to design effects in clustered survey sample designs. Strong conditional dependence can render five distinct ratings of aspects of a performance no more informative than a single rating would have been. Conversely, when conditional dependence effects are weak, the five ratings from the same task provide almost as much information as observations from five distinct tasks. One way to model dependence is by introducing a parent that affects only responses in Task *j*, like the one labeled in Figure 10 as $\text{TaskContext}_j$. A high value for $\text{TaskContext}_j$ shifts probabilities up a bit for each

observable from this task, beyond the probabilities produced by $\theta$, a low value shifts all the probabilities down a bit, by amounts that can be estimated from response data (Bradlow, Wainer, & Wang, 1999).

Figure 11 shows posterior probabilities based on a student's responses to four complex tasks, one each from Task Families 1, 3a, 3b, and 3c. For example, the first task is from the 3a family and has two observables, *vocabulary* and *explicit meanings.* Vocabulary has the SM variable *writing* as a parent, explicit meanings has both *writing* and *reading,* and both observables have this task's TaskContext variable as a parent to allow for conditional dependence. The distributions reflect belief after observing a response pattern that might be produced by a student who has high

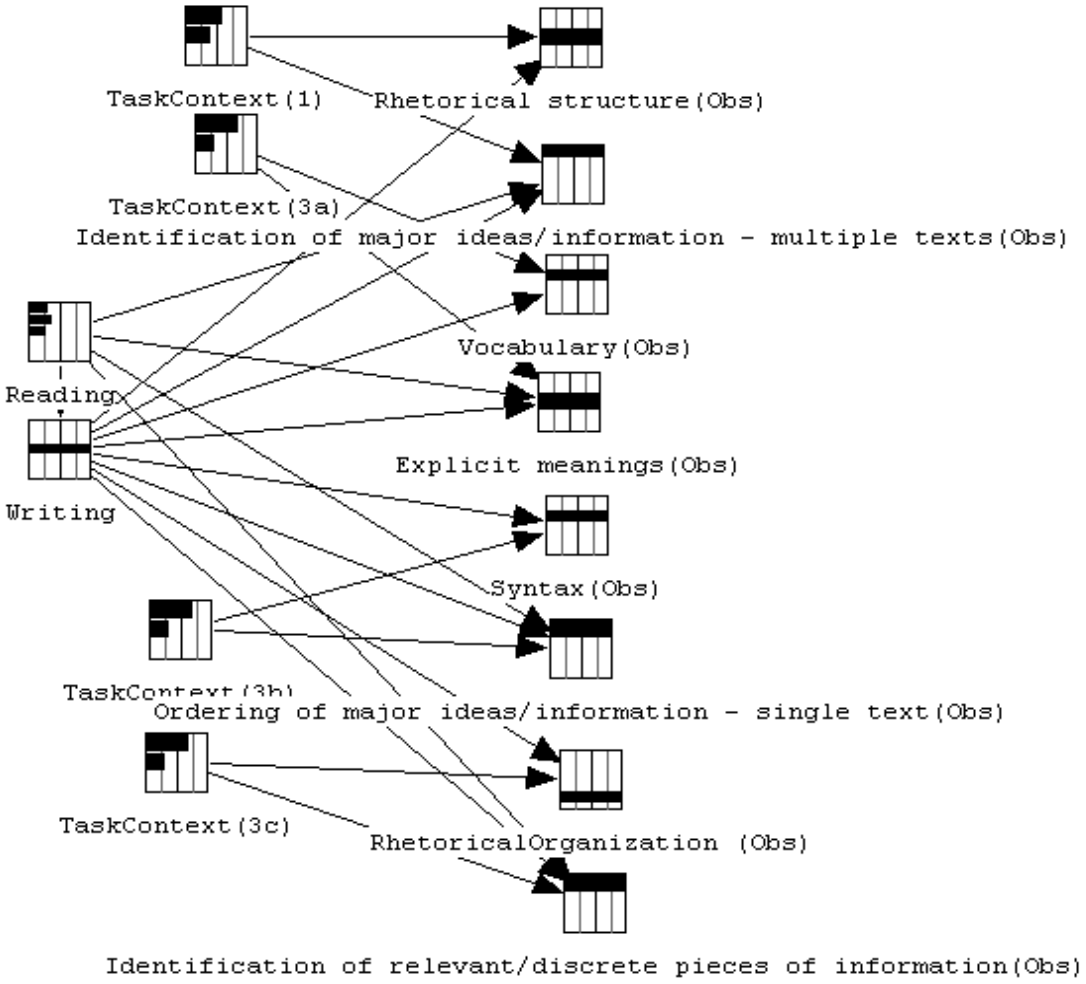



*Figure 11.* Updating probabilities under the many-to-many model for four tasks.

proficiency in reading but moderate proficiency in writing (i.e., just above the hurdle to produce evidence about comprehension). The realized values of the observables are indicated by probability bars that go all the way to one for the observed values, for they are now known with certainty. The status of the reading and writing SM variables reflects evidence accumulated over all observables for all four tasks, reasoning backwards via Bayes Theorem through conditional probability matrices. The patterns in the conditional probability matrices were determined by the features of the tasks and the stress they put upon different aspects of proficiency. Accordingly, for reading, we see a posterior distribution with most of the probability across the higher ranges of proficiency; for writing, we see a concentration of probability in the middle of the range.

**The Assembly Model**

Student, evidence, and task models define a universe of tasks an examinee might be presented with, procedures for evaluating what is observed, and machinery for updating beliefs about the values of the student-model variables. *Assembly specifications* on the assembly model define the mix of tasks that will constitute a given student's assessment. A content-by-process matrix is a familiar form that some assessment assembly specifications take. One can impose further constraints that concern statistical characteristics of items, in order to increase measurement precision, or that concern non-statistical considerations such as content, format, timing, and cross-item dependencies (Berger & Veerkamp, 1996). By managing these considerations, the assembly model provides a mechanism for (a) setting the range of circumstances that are covered for providing evidence about a student, (b) controlling the difficulty of tasks not only overall but with respect to aspects of competence or capabilities in various performance settings, (c) managing which information tends to accumulate in the form of distributions for SM variables and what information does not accumulate and thus constitutes noise in the statistical model (Almond & Mislevy, 1999), and (d) establishing the amount of evidence that is accumulated about the various student-model variables, and hence the claims, through the number and nature of observations that bear on them (i.e., reliability). The assembly model is where we determine "what accumulates"—that is, how we effect the meaning of SM variables.

**What Accumulates?**

Performance on any task, large or small, entails many aspects of students' knowledge. A work product from a given task has the potential to provide evidence about any of those aspects separately or about various amalgamations of them. But simply observing a performance does not constitute "measuring" anything. Both components of the evidence model represent necessary steps between capturing work products and measuring students' capabilities. The answer to "What does this task measure?" is "By itself, nothing yet." Just which claims will be informed by performances on that task, which aspects of knowledge or capabilities will be measured, and at what grainsize, will be determined by

- the features of performance that are captured as work product(s),
- the features of the work product(s) that are specified in the evaluation component of the evidence model,
- the knowledge and skill demands of other tasks (managed in terms of values of task-model variables) called for in assembled assessments, and
- the student-model variables, in terms of the patterns in measurement models by which observables are modeled as functions of the SM variables.

The idea of accumulating evidence is fundamental to measurement (Green, 1978). While every task calls upon a unique mixture of knowledge and skill, the mixtures tapped by two tasks will differ in some ways more than others. Similarities in behaviors in those situations are attributed to commonalities in the knowledge they demand. A total test score, for example, has accumulated bits of information across many tasks, reflecting whatever it is they have in common. Whatever knowledge one item requires that others do not has decreasing influence on the score as test length increases. Classical test theory formalizes this idea by quantifying how much information items share for a given group of examinees. The additional structure of item response theory further connects a person's score with expected performance on each item. In either case, the more examinees differ from one another on the knowledge that does not accumulate, the less accurate are the scores over the same number of items.

A cardinal principle of our view of assessment design is that what accumulates over tasks should be intentional rather than accidental. This is accomplished by identifying features of tasks that call upon pre-identified aspects of knowledge in

various ways and in varying degrees. For instance, we mentioned earlier the strategy of measuring multiple components of language competence individually, with separate tests. Constructing the subtest for each component proceeds by using instances of task models that have features that call for that knowledge or skill at levels of interest, and features that call upon other aspects at levels that should not pose problems to the examinees of interest. With multivariate student models, we influence which tasks contribute evidence about which SM variables by managing the key features of individual tasks in terms of values in the task models according to which they are generated.

### The Presentation Model

The Presentation Model specifies how the assessment elements framed in the design models (*student*, *evidence*, and *task models*) will actually look and operate within a particular delivery environment—for example, how tasks are to be presented, how evidence is to be identified and accumulated, how tasks are to be scheduled. It defines the grammar and the semantics, but not the specific content, of messages that link the processes and manage the flow of control through the operational assessment system. For the purposes of the present paper, we don't need to say much more than that the presentation model provides specifications for how student, evidence, task and assembly models are operationalized in the appropriate processes of the delivery system. The presentation model in the CAF is to be distinguished from the *presentation process* in an implemented assessment. The Presentation Model contains specifications, logic, and syntax that the operational components in an implemented assessment, whereas the presentation process administers an assessment in accordance with those specifications. An advantage of defining a Presentation Model, rather than just building the processes, is that it becomes easier to task presentation processes out to external vendors or different groups of implementers.

## A Four-Process Delivery System

This section discusses the framework for assessment delivery systems, at a level sufficient to complete the chain from substantive argument to operational elements. The reader is referred to Almond et al. (2001) for more on the four-process delivery system.

The bottom panel of Figure 1 depicts the four principal processes that take place in assessment delivery, and Table 6 summarizes their functions and the messages (data structures) that are passed among them. Some of the processes and messages are collapsed or hidden in familiar forms of assessment. In informal assessments such as conversations between students and teachers, none of the

Table 6

Summary of the Four-Process Assessment Delivery System

| Name of process | Functions | Input | Output |
|---|---|---|---|
| Activity selection | Determine what to do next: Select task, stop testing, branch to instruction, etc. | From Administrator: Task selection algorithm & rules.<br>From T/E: Descriptive data about available tasks.<br>From EAP: Current belief about student. | To PP: Identification of task(s) to present. |
| Presentation | Manage presentation of task to student: Present stimulus material, provide tools and affordances, capture work product. | From AS: Identification of task(s) to present.<br>From T/E: Presentation material, location of tools, specs for activity management within task.<br>From student: Actions, product(s). | To student: Task materials & affordances.<br>To EIP: Work product(s). |
| Evidence identification (task-level scoring) | Process evidence from individual tasks: Extract and summarize salient aspects of students' performance, provide task-level feedback. | From PP: Work product(s).<br>From T/E: Rules & data for evaluating WP. | To EAP: Values of observable variables for this task.<br>To student/other users: Task-level feedback. |
| Evidence accumulation (test-level scoring) | Integrate evidence across tasks: Maintain distribution for SM variables of each student, update as new evidence arrives. | From EIP: Values of observable variables.<br>From T/E: Functions & parameters ("weights of evidence") for synthesizing evidence from this task into scoring model. | To ASP: Current belief about student.<br>To student/other users: Test-level feedback. |

*Note:* Abbreviations: WP = Work Product, T/E = Task/Evidence Composite Library, ASP = Activity Selection Process, PP = Presentation Process, EIP = Evidence Identification Process, and EAP = Evidence Accumulation Process.

processes is explicit or delineated, but it is possible to analyze the interactions in these terms. This depiction makes no assumptions about use of any particular platform (e.g., human, paper, computer) for any of the processes—any or all are possible in differing combinations, depending on the purposes and constraints of a particular assessment. Further, the processes can be arranged to communicate with each other in many different patterns, over time and across space, and also depending on assessment purposes and resources. We can describe, as particular different forms and patterns of interaction among these same general processes, assessments as different as intelligent tutoring systems, large-scale remotely scored tests like the SAT, computerized adaptive tests, self-guided practice workbooks, negotiated assessments like dissertations, and accumulations of work over time like the AP Studio Art Portfolio Assessment.

Sitting in the center and informing all four processes is the *task/evidence composite library* (for short, the *task library*). It contains task materials and the information needed to select them, present them, score them, and use the results to update beliefs about students. Each of these task/evidence composites is constructed and linked in accordance with task and evidence models. The task/evidence composite, therefore, contains the information that each of the processes requires in order to produce its own assessment data structures.

The *activity selection process* selects a task or a collection of tasks and instructs the presentation process to display it. When the examinee has finished interacting with a specified set of tasks (maybe just one, maybe a predefined group, maybe an entire test form), the presentation process sends the results (a possibly complex and multi-part *work product*) to the *evidence identification process*. This process evaluates observable variables and can do two things with them: It can present task-level feedback based on some or all of them, and it can pass some or all of them to the *evidence accumulation process*. The evidence accumulation process updates the *scoring model*, a copy of the CAF student model that has been individualized to that examinee. If the process has multiple phases, the activity selection process again decides what to do next. For different purposes, different people play the roles of the various actors in the system (test administrator, examinee, user of immediate feedback, user of summary reports), and the same person can play several roles. In a coached practice system, for example, the student plays the role not only of the examinee, but also of a user of local feedback and also of a test administrator when the student decides which tasks to work on next.

These processes are common to all assessments. To maximize flexibility and achieve efficiencies in building assessments to accommodate a range of purposes and domains, one may implement the processes as physically discrete as well as logically discrete. The purpose of an assessment drives the sequencing of the flow of control among them as well as how they are realized, as we will see in the language assessment examples. If processes have physical as well as logical integrity, they can be re-used in different combinations for different purposes, avoiding the development of inflexible delivery systems whose purposes are "hard-wired in" to specific applications.

What is the relationship between the objects in the CAF and the processes in the assessment cycle? We see in Table 6 that each process needs to receive certain information from other processes and the task/evidence library does something with that information and passes the results on to other processes. The information passed on by the presentation process is the work product, for example, in the form that was specified in the task model and is anticipated by the evaluation rules in the evidence model (e.g., rubrics, scoring keys, evaluation algorithms). The models of the CAF are where the nature, the structure, and the flow of this activity have been explicated, in a way that operationalizes the substantive argument.

From an operational perspective, the CAF models describe the properties and specifications for all of the information passing among processes. From the perspective of implementation, the CAF models are interlocking schemas, copies of which will be filled in with specifics for every task, every student, and every statistical distribution in the operational assessment. From a conceptual perspective, the CAF models provide blueprints for the physical form and operational procedures needed to acquire data and carry out reasoning within the structure of the evidentiary argument. The substantive argument is not apparent in the particulars of tasks and responses and variables once the assessment has been implemented and the processes are running, but it is the justification for the inferences that are drawn from the results of that activity. The key connections are suggested in Figure 1 by dashed lines from particular CAF models to particular processes.

- The student model lays out the form of the operational scoring models and thus the information that will be available for activity selection and summary reporting. Summary reporting consists of functions of posterior distributions for SM variables, be they numerical, graphical, or mixed

verbal and quantitative, along with an indication of amount of evidence for substantively stated claims (e.g., posterior standard deviations, probability of mastery, etc.).

- The task model provides formats for presentation material and work products used by the presentation process and the evidence identification process, and, in terms of TM variables, description data about tasks that the activity selection process can use for task selection. It also describes data structures for task level feedback and evaluation rule data.

- From the evaluation component of the evidence model, the evaluation rules describe the procedures that the evidence identification process needs to carry out, acting on the work product, to produce values of observable variables.

- The models in the Measurement Component of the evidence model provide the structure and specifications for task parameters that the evidence accumulation process needs to update the SM variables in examinees' scoring models in light of the values of their observable variables.

- The assembly model describes the strategy used for selecting tasks, which are implemented by the activity selection process. These strategies can depend on information in the current *examinee record* (described in the student model) and use descriptive data about the task (described in terms of TM variables in the task model).

Working through the CAF models accomplishes the necessary coordination among the activities carried out in the disparate processes, often by different people at different times and places. Suppose we want to use a given measurement model to update a scoring model in light of data produced from a task written under a given task model (TM). Before the assessment is implemented, it will have been ascertained that the evaluation component of the evidence model (EM) conforms with that TM as to the specifications for work products; that the evaluation component of the EM conforms with the measurement component as to the observable variables; and that the measurement component conforms with the student model as to the knowledge, skill, and ability variables.

**Example (Cont.): Delivery Systems for the Language Assessment Examples**

This section sketches hypothetical language assessments built to serve Purposes 1-4. The examples illustrate rather different configurations of delivery processes to show how flexibly the delivery processes can be re-organized; to show that the same assessment design framework informs assessments of very different

kinds; and to indicate the relationships in delivery systems that remain constant even as the processes are organized in different flows for different purposes.

Purpose 1 is a high-stakes proficiency test for students in all departments, and many students are tested at the same time during orientation. The flow of activity is one loop around the diagram, starting with activity selection. All students receive the same test form at once, so there is no tailoring or randomization of tasks to result in different students taking different assessments. The presentation process is paper-and-pencil based, using standardized administration conditions in the gymnasium. Students write three essays by hand in a blue book; the essays are the work product that will be evaluated. The books are sent to teams of raters who assign to each essay a holistic rating of overall quality. Each essay is read by two raters. The raters use light pens and bar codes to enter the scores they assign into one large file, the format of which was also specified in the evidence model. The evidence accumulation process uses a rater-effect IRT model to accumulate evidence across essays and later check for anomalous ratings. The students' IRT $\theta$ estimates are sent to the English department, where a pass/fail cut point is determined and pass/fail results are returned to students.

Purpose 2 is also a high-stakes test, which uses tasks of the same type except that they require knowledge about psychology. Despite the similarity of the tasks, however, the delivery system flow differs in several respects. The test is administered in the supervised psychology lab over the campus intranet, at a time of each student's choice. Activity selection is interactive: From the many essay topics that are available, the student chooses the first, and the second is selected at random. The presentation process is implemented on the intranet. After each topic has been selected, the student keys in an essay. There is a two-part work product, namely the contents of the two essays. It is emailed to Prof. Jones who carries out the evidence identification process in accordance with the scoring rubrics and previous years' examples. He assigns a single rating to each essay. The evidence accumulation process uses no measurement model and thus provides no estimate of accuracy: Professor Jones simply adds the two scores together, compares them with a cut score, and emails each student a pass/fail result.

Purpose 3 is addressed in the university's language lab, with the intention of assigning students to classes that will develop their academic reading and writing proficiencies. Only students who did not pass the Purpose 1 assessment are tested, so it is known that they do not have the overall proficiency that encompasses

sufficiently high levels of both academic reading and writing skills in English. The activity selection process is random within strata defined by *task families* that taken together, provide evidence across the range of reading and writing proficiencies. In particular, each student writes responses to three more prompts, one each selected at random by the lab director from task families 3a, 3b, and 3c. Regarding the presentation process: Hand-written responses in blue books contain the work products, again written responses but simple written responses for the first two tasks and an essay from just the third. These work products are forwarded to graduate assistants who carry out the evidence identification process; they rate each response with respect to the several analytic rating scales described in the section on *evaluation rules.* Which scales are used for which kinds of tasks is indicated to the grad assistants, part of the evidence rule data. A vector of values for observable variables is created for each task response on the rating scales that apply to that task. These data are passed along to evidence accumulation process, which utilizes Prof. Roberts' multivariate latent class model. The student model consists of Master / Nonmaster with respect to the content of each of six instructional modules at the language lab, which have different balances of reading and writing emphasis. Prof. Roberts has estimated conditional probabilities for ratings in the different scales on each of the tasks in the pool, and can calculate posterior probabilities of mastery for each module based on a student's vector of values for the observable variables across the three tasks. Rated aspects of reading, writing, and combinations of both are thus synthesized into this profile. The summary report is a page of verbal descriptions of the student's strengths and weaknesses in terms that are related to the instructional goals of the modules, to be used by the student and an advisor in planning a course of study.

Purpose 4 is a computer-based practice program, designed to help students prepare for the Purpose 1 test. Although the stimulus materials and instructions for the tasks are identical to those used in the Purpose 1 tests from previous years, the flow in the delivery system is different because the assessment has different objectives. Activity selection is wholly controlled by the student. The student can choose to work on any task in the library, at any time, in any order, and as many times as the student likes. The presentation process is managed by the computer program; it provides the student access to the reading materials that are required, and provides the same computer-based essay-writing tool the Psychology department uses for Purpose 2 to write essays (i.e., to create work products). The

student can choose to have an essay scored or not and can proceed back to take a different prompt if desired. These are decision-making processes within activity selection that the student carries out in real time. If the student does decide to have an essay scored, the essay is passed to the evidence identification process that resides on the same computer. Two phases of automated evaluation rules are applied. The first is parsing of the essay for lexical and grammatical features of the response. The second is entering these extracted features into a neural network scoring algorithm, which was calibrated with human raters' evaluations of essays written to the same prompts and stimulus material when they were used for Purpose 1 decisions in previous years. The resulting scores are values of observable variables, the usual output of evidence identification. These are reported directly and immediately to the student as task level feedback. No evidence accumulation is built into the system, although the student or her coach may do some informal evidence accumulation of their own outside the system to monitor the student's progress over time.

## Closing Comments

We have endeavored to lay out structural relationships that underlie coherent assessments—relationships among assessment purposes, substantive experience and theory, statistical models and task authoring schemas, and the elements and processes of operational assessments. We have placed some emphasis on measurement models that integrate multivariate student models, complex assessment tasks, and task features that mediate the relationships, because this region of the structure is rare in current practice.

While the ECD framework is useful for both analyzing existing assessments and designing new ones, it is the latter that should prove more immediately useful. This is particularly the case for assessments that aim to take advantage of advances in technology (how to capture data, produce interactive simulations, and do assessment over the internet) and psychology (concerning the nature of knowledge and the social and cognitive structures and processes by which people acquire and use knowledge). Neither advances in gathering data nor understanding cognition are sufficient for improving the practice of assessment, however. When substantive insights, technological advances, or new uses are grafted onto existing assessment systems, mismatches are inevitable. Ad hoc and intuitive reasoning from new kinds of data for different purposes may be a proper start, but they are poorly suited for

designing large-scale and high-stakes assessments. For these circumstances, the quantitative framework for reverse reasoning that probability-based reasoning offers is unparalleled. The intricacies of our language assessment examples suggest just how unlikely it is for all of the interconnections among the multivariate measurement models and the complex assessment tasks to come together without a framework that coordinates everyone's work—let alone for it to happen efficiently, and in terms of re-usable design objects and operational processes.

When unexamined standard operating procedures fall short, it is often worth the effort to return to first principles as we have done here—to formal analyses, explicit structures, and normative models. Initial applications of these ideas may be labor intensive and time consuming. However, practice will advance not just from presentations like this one, but from working examples, re-usable elements, and pieces of infrastructure from initial applications, all of which can be adapted to new projects. That there is a layer of formalization and principles that underlies this work is essential; that it be familiar in detail to all those who employ the ensuing processes and structures is not. An explicit conceptual framework that can be shared, debated, and improved will speed the diffusion of improved practices in assessment and will facilitate the development of exemplars and infrastructure. It is to this end we hope the present paper will contribute.

## References


Adams, R., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Almond, R. G., DiBello, L., Jenkins, F., Senturk, D., Steinberg, L., & Yan, D. (2001, January). *Models for conditional probability tables in educational assessment.* Paper presented at AI and STATISTICS 2001, the Eighth International Workshop on Artificial Intelligence and Statistics, Key West, Florida.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223-237.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2001). *A sample assessment using the four process framework* (CSE Tech. Rep. No. 543). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation. http://www.cse.ucla.edu/CRESST/Reports/TECH543.pdf

American Council on the Teaching of Foreign Languages (1999). *ACTFL proficiency guidelines: Speaking (Revised 1999).* Hastings-on-Hudson: Author.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, England: Oxford University Press.

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

Berger, M. P. F., & Veerkamp, W. J. J. (1996). A review of selection methods for optimal test design. In G. Engelhard, & M. Wilson (Eds.), *Objective measurement: Theory into practice (Vol. 3).* Norwood, NJ: Ablex.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

Brennan, R. L. (1983). *The elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Brindley, G. (1994). Task-centered assessment in language learning: The promise and the challenge. In N. Bird, P. Falvey, A. Tsui, D. Allison, & A. McNeill (Eds.), *Language and learning: Papers presented at the Annual International Language in Education Conference, Hong Kong, 1993* (pp. 73-94). Hong Kong, China: Hong Kong Education Department.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

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.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Urbana, IL: University of Illinois Press.

Dibello, L.V., Stout, W.F., & Roussos, L.A. (1995). Unified cognitive / psychometric diagnostic assessment likelihood based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-389). Hillsdale, NJ: Erlbaum.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

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

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.

Ericcson, K. A., & Smith, J. (1991). *Toward a general theory of expertise.* Cambridge, England: Cambridge University Press.

Falmagne, J.-C., & Doignon, J.-P. (1988). A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology, 41,* 1-23.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Frase, L. T., Chudorow, M., Almond, R. G., Burstein, J., Kukich, K., Mislevy, R. J., et al. (in press). Technology and assessment. In H. F. O'Neil & R. Perez (Eds.), *Technology applications in assessment: A learning view*. Hillsdale, NJ: Erlbaum.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman and Hall.

Green, B. (1978). In defense of measurement. *American Psychologist, 33,* 664-670.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1997). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-47). New York: Simon & Schuster Macmillan.

Habermas, J. (1971). *Knowledge and human interests.* (Translated by Jeremy Shapiro). Boston: Beacon Press.

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

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

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults.* Princeton, NJ: National Assessment of Educational Progress / Educational Testing Service.

Linacre, J. M. (1989). *Multi-faceted Rasch measurement.* Chicago: MESA Press.

Martin, J. D., & VanLehn, K. (1995). A Bayesian approach to cognitive assessment. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 141-165). Hillsdale, NJ: Erlbaum.

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. *Education Researcher, 32*(2), 13-23.

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

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 437-446). San Francisco: Morgan Kaufmann.

Mislevy, R. J., Senturk, D., Almond, R. G., Dibello, L. V., Jenkins, F., Steinberg, L. S., et al. (2002). *Modeling conditional probabilities in complex educational assessments* (CSE Tech. Rep. No. 580) Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Available from http://www.cresst.org/summary/580.htm

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (in press). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice.* Hillsdale, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (in press). Making sense of data from complex assessment. *Applied Measurement in Education*. Available from http://www.cse.ucla.edu/CRESST/Reports/RML%20TR%20538.pdf

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (Center for Performance Assessment Research Report). Princeton, NJ: Educational Testing Service.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*(4), 342-366.

Pellegrino, J., Chudowsky, N., and Glaser, R., (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Research Council's Committee on the Foundations of Assessment. Washington, DC: National Academy Press.

Pirolli, P., & Wilson, M. (1998). A theory of the measurement of knowledge content, access, and learning. *Psychological Review 105*(1), 58-82.

Rasch, G. (1960 / 1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research / Chicago: University of Chicago Press (reprint).

Reckase, M. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Sarbin, T. R., Taft, R., & Bailey, D. E. (1960). *Clinical inference and cognitive theory*. New York: Holt, Rinehart, and Wilson.

Scheiblechner, H. (1972). Das lernen und lösen komplexer denkaufgaben (The learning and solution of complex cognitive tasks). *Zeitschrift für experimentalle und Angewandte Psychologie, 19*, 476-506.

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

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

Sheehan, K. M., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models in a measure of document literacy. *Journal of Educational Measurement, 27*, 255-272.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., & Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statistical Science, 8*, 219-283.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto, (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

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

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco: Jossey-Bass.