On the Roles of Task Model Variables in Assessment Design

CSE Technical Report 500

Robert J. Mislevy, Linda S. Steinberg, and Russell G. Almond
Educational Testing Service, Princeton, New Jersey

January 1999

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.2 Validity of Interpretations and Reporting of Results - Evidence and Inference in Assessment Robert J. Mislevy, Project Director, CRESST/ Educational Testing Service

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U. S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U. S. Department of Education.

ON THE ROLES OF TASK MODEL VARIABLES IN ASSESSMENT DESIGN[1]

Robert J. Mislevy, Linda S. Steinberg, and Russell G. Almond

Educational Testing Service, Princeton, New Jersey

ABSTRACT

Tasks are the most visible element in an educational assessment. Their purpose, however, is to provide evidence about targets of inference that cannot be directly seen at all: what examinees know and can do, more broadly conceived than can be observed in the context of any particular set of tasks. This paper concerns issues in assessment design that must be addressed for assessment tasks to serve this purpose effectively and efficiently. The first part of the paper describes a conceptual framework for assessment design, which includes models for tasks. Corresponding models appear for other aspects of an assessment, in the form of a student model, evidence models, an assembly model, a simulator/presentation model, and an interface/environment model. Coherent design requires that these models be coordinated to serve the assessment's purpose. The second part of the paper focuses attention on the task model. It discusses the several roles that task model variables play to achieve the needed coordination in the design phase of an assessment, and to structure task creation and inference in the operational phase.

[1] This paper was presented at the conference "Generating items for cognitive tests: Theory and practice," co-sponsored by Educational Testing Service and the United States Air Force Laboratory and held at the Henry Chauncey Conference Center, Educational Testing Service, Princeton, NJ, November 5-6, 1998.

> A construct-centered approach [to assessment design] would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (Messick, 1992, p. 17)

INTRODUCTION

Task Design and Assessment Design

Tasks are properly a central focus of educational assessment, because they produce the evidence upon which any subsequent feedback, decisions, predictions, or placements are based. Historically, task design has been regarded more as an art than a science. Today, however, pressures from several directions impel us to consider more principled approaches to task design. The economies of continuous large-scale computerized testing, for example, consume far more test items than testing at limited occasions. Research from cognitive and educational psychology is providing insights into the structure and acquisition of knowledge, and offering clues about alternative ways that knowledge can be evidenced. New technologies for simulating work environments beg the question of how to make sense of the rich data they can produce.

An assessment designer in this new world will have to create tasks with a credible argument for how students' behaviors constitute evidence about targeted aspects of proficiency, and a clear structure for how the tasks will be produced, presented, and evaluated. Task design is thus an element of assessment design more broadly conceived. The present paper discusses task design in this light. We borrow terminology and concepts from our *evidence-centered* assessment design project, Portal, to frame the discussion. The ideas are illustrated with examples from two assessments: The Graduate Record Examination (GRE), as a prototypical large-scale standardized assessment, and HYDRIVE, a coached practice system, as a representative product constructed explicitly from a cognitive perspective and using simulation technologies. The point is to see these two seemingly quite different assessments as instantiations of the same underlying design elements.

Evidence-Centered Assessment Design

When the toolkit of standard testing practice lacks off-the-shelf procedures to develop an assessment that takes advantage of new technologies or builds around an alternative psychological theory, the designer must return to first principles. To this end, assessment design can profitably be considered from the perspective of evidentiary reasoning as it is developed in the work of David Schum (1987, 1994). Schum draws upon themes and tools from centuries of scholarly research, in fields that range from philosophy and jurisprudence to statistics and expert systems. He argues that while every realm of human activity has evolved its own specialized methods for evidentiary reasoning, common underlying principles and structures can be identified to improve applied work in all of them. These foundational principles of evidentiary reasoning are especially useful for attacking new or novel problems, when standard solutions and familiar methods fall short.

Our objective is to exploit this perspective, and the principles and tools gained thereby, in the domain of educational assessment. We use the term *assessment* broadly, to include not only large-scale standardized examinations but classroom tests both formative and summative, coached practice systems and intelligent tutoring systems, even conversations between a student and a human tutor. All face the same essential problem: drawing inferences about what a student knows, can do, or has accomplished, from limited observations of what a student says or does. An evidentiary perspective helps sort out the relationships among what we want to infer about examinees, what we can observe that will provide evidence to back our inferences, and situations that enable us to evoke that evidence.

Evidence-centered assessment design squares well with Messick's (1992) construct-centered approach, epitomized in our introductory quote. The difference is mainly a matter of emphasis. Messick accents the importance of conceptualizing the target of inference, or just what it is about students the assessment is meant to inform. As Yogi Berra said, "If you don't know where you're going, you might end up someplace else." We stress the stages of acquiring and reasoning from evidence, because the field lacks off-the-shelf methodologies for structuring inference with the richer data and more complex student models

that are now beginning to appear in educational assessment. Either way, the key ideas are these:

Identifying the aspects of skill and knowledge about which inferences are desired. A given assessment system is meant to support inferences for some purpose, whether it be course placement, diagnostic feedback, administrative accountability, guidance for further instruction, licensing or admissions decisions, or some combination of these. In order to support a given purpose, how should we characterize examinees' knowledge?[2]

Identifying the relationships between targeted knowledge and behaviors in situations that call for their use. What are the essential characteristics of behavior or performance that demonstrate the knowledge and skills in which we are interested? What do we see in the real world that seems to distinguish people at different levels of proficiency in these respects?

Identifying features of situations that can evoke behavior that provides evidence about the targeted knowledge. What kinds of tasks or situations can elicit the behaviors or performances that demonstrate proficiency? The way we construe knowledge and what we consider evidence about it should guide how we construct tasks and evaluate outcomes.

The objective of the Portal project is to create a conceptual framework and supporting software tools for designing assessments in this light. The project has three distinguishable aspects: (a) An evidence-centered perspective on assessment design; (b) object definitions and data structures for assessment elements and their interrelationships; and (c) integrated software tools to support design and implementation. In this paper we draw upon the perspective and a high-level description of the central objects and interrelationships. In particular we will explore aspects of the Portal *task model*. We draw out connections between features of tasks and various assessment functions including task construction, inference, reporting, and validity argumentation - all of which can be described in terms of the roles of task model variables.

[2] We use the term "knowledge" broadly, to encompass its declarative, strategic, and procedural aspects, and recognize that a person's knowledge is intertwined with social, cultural, and technological contexts. This latter understanding is central to argument from evidence to implication, and as such, equally critical to assessment design and validity investigations.

The following section sets the stage for this discussion by laying out the essential structure of the Portal *conceptual assessment framework*. Following that, we consider the various and interconnected roles of task model variables.

A MODEL FOR EVIDENCE-CENTERED ASSESSMENT DESIGN

Overview of the Basic Models

Figure 1 is a schematic representation of the six highest-level objects, or models, in a Portal conceptual assessment framework (CAF). These models must be present, and must be coordinated, to achieve a coherent assessment. We would claim that these basic models are present, at least implicitly, and coordinated, at least functionally, in existing assessments that have evolved to serve well some inferential function. Making this structure explicit helps an assessment designer organize the issues that must be addressed in creating a new assessment. Retrospectively, it helps clarify how pervasive design issues have been managed in successful assessments in the past, or overlooked in failures.

These are the basic models:

The *Student Model* contains variables representing the aspects of proficiency that are the targets of inference in the assessment, and it is where we manage our uncertain knowledge about these variables. Student model variables thus concern characteristics of *students*.

The *Evidence Model* describes how to extract the key items of evidence (values of *observable variables*) from what a student says or does in the context of a task (the *work product*), and models the relationship of these observable variables to student-model variables. Observable variables concern characteristics of *performances*.

The *Task Model* describes the features of a task that need to be specified when a task is created. We will use the term *task* in the sense proposed by Haertel and Wiley (1993), to refer to a "goal-directed human activity to be pursued in a specified manner, context, or circumstance." A task can thus include an open-ended problem in a computerized simulation, a long-term project such as a term paper, an language-proficiency interview about an examinee's family, or a familiar multiple-choice or short-answer question. We

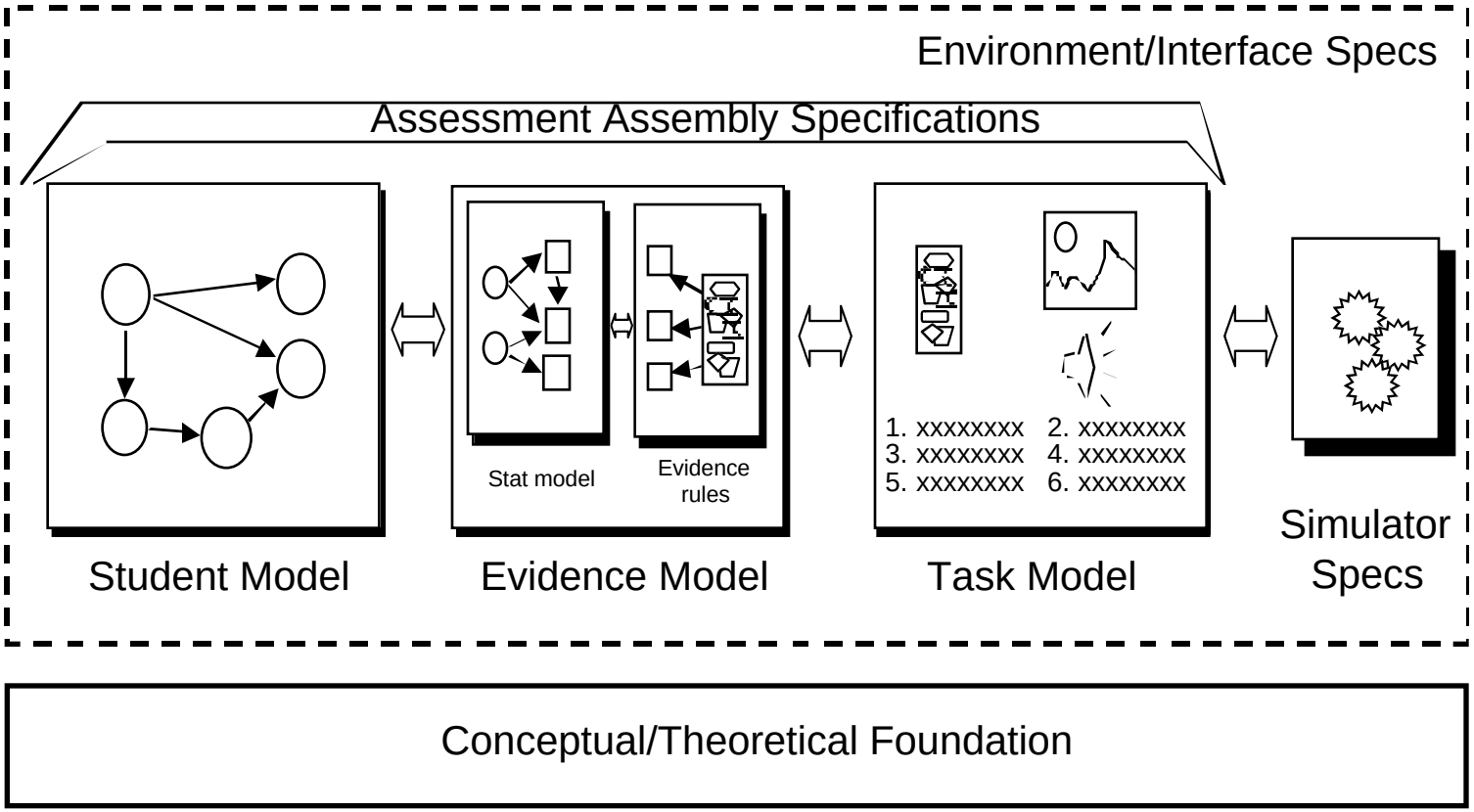


Figure 1. High-level objects in the Conceptual Assessment Framework

will reserve the term "item" for these latter cases. Task model variables concern characteristics of the *situations* by which evidence is obtained.

The *Assembly Model* describes the mixture of tasks that go into an operational assessment, or the procedure for determining tasks dynamically if appropriate.

The *Simulator Model* describes the environment in which a particular task will run and the capabilities that are required for this to happen. The term could refer literally to a computerized simulation environment, but more generally construed can refer to the familiar, non-interactive, presentation of items in a paper-and-pencil (P&P) examination or to the procedure through which a human interlocutor conducts an interview to gather evidence about an examinee's language proficiencies.

The *Environment Model* describes the overall assessment environment. This includes specifications for whatever is needed to carry out the assessment, such as physical requirements, tools for examinees, computer hardware and software, timing requirements, security procedures, and so on.

In general, a CAF will have one operational student model, assembly model, and environment model. It may, however, have many different operational evidence, task, and simulator models.

The Student Model

The student model directly answers Messick's question, "What complex of knowledge, skills, or other attributes should be assessed?" Student-model variables describe characteristics of examinees (knowledge, skills, abilities) about which the user of the assessment wants to make inferences (decisions, reports, diagnostic feedback, advice).

Configurations of values of student-model variables are meant to approximate selected aspects of the countless skill and knowledge configurations of real students, from some perspective about how to think about skill and knowledge. This is how we want to talk about the student, although we can't observe the values directly. There could be one or hundreds of variables in a student model. They could be qualitative or numerical. They might concern tendencies in behavior, use of strategies, or ability to apply the big ideas in a

domain. The factors that determine the number and the nature of the student model variables in a particular application are the conception of competence in the domain and the intended use of the assessment.[3] A test used only for selection, for example, might have just one student-model variable, overall proficiency in the domain of tasks, while a diagnostic test for the same domain would have more student-model variables, defined at a finer grain-size and keyed to instructional options.

Defining student model variables specifies our target(s) of inference. At the beginning of a given student's assessment, we know little about the values of this student's variables and wish to sharpen our knowledge. We move from a state of greater uncertainty to lesser uncertainty about these unknown values by making observations which provide evidence about them, and integrating this new information into our beliefs.

In Portal, we use Bayesian inference networks, or Bayes nets for short (Jensen, 1996), to manage our uncertain knowledge about the student. The student model is a fragment of a Bayes net, the student model variables being the variables in the network. A joint probability distribution for these variables at a given point in time represents our knowledge about the values of the student-model variables corresponding to a particular examinee. We update this distribution when we make an observation. (We'll say a bit more about how this is done in the following section on Evidence Models, but see Mislevy & Gitomer, 1996, for a more complete discussion.) Belief about a given person's values before an assessment could be based on background information or prior experience with that examinee. Uninformative distributions would usually be used as the prior distribution for all examinees in a high-stakes test, though, because considerations of fairness demand that only information from the assessment at hand enter into the summary of their performances.

Example 1. Figure 2 graphically depicts the student model that underlies most familiar assessments: a single variable, typically denoted θ, that represents proficiency in a specified domain of tasks. We use as examples the paper and pencil (P&P) and the computer adaptive (CAT) versions of

[3] As we shall see, task model variables play a central role in defining student model variables operationally. The idea is for this operational definition to be the result of purposeful planning, rather than an coincidental outcome of task creation.

Figure 2. The student model in the GRE Verbal measure contains just one variable: the IRT ability parameter θ, which represents the tendency to make correct responses in the mix of items presented in a GRE-V.

the Graduate Record Examination (GRE), which consist of domains of items for Verbal, Quantitative, and Analytic reasoning skills. Our knowledge before the test starts is expressed as an uninformative prior distribution. We will update it in accordance with behaviors we see the examinee make in various situations we have structured; that is, when we see her responses to some GRE Verbal test items.[4]

Example 2. Figure 3 is a more complex example of a student model, taken from Gitomer and Mislevy (1996). It is based on HYDRIVE (Steinberg & Gitomer, 1996), a coached practice system that ETS built to help Air Force trainees learn to troubleshoot the hydraulic systems of the F15 aircraft. Students worked their way through problems in a computer simulated environment much as they would on the flight line. The variables of the student model were used to capture regularities in the student's behavior, and their tendencies to use identified expert troubleshooting strategies.

This student model is a fragment of a Bayes net, and these nodes are the student model variables. Student-model variables were derived in light of cognitive task analyses (CTA) of the job, the purpose of the HYDRIVE system, and the instructional approach of the system. The CTA showed

[4] The simplicity of this student model is deceptive, by the way. It takes a great deal of hard work to make such a simple model work well. In order to be appropriate for capturing and expressing information from potentially thousands of different items, some very sophisticated interrelationships are posited. Much care is taken in just how and which are to be observed for a given examinee; empirical evidence is be carefully checked to avoid inferential errors that lead to certain kinds of unfair inferences. In the second half of the paper we will mention some of the considerations that are needed to ensure this simple model will suffice in the GRE.

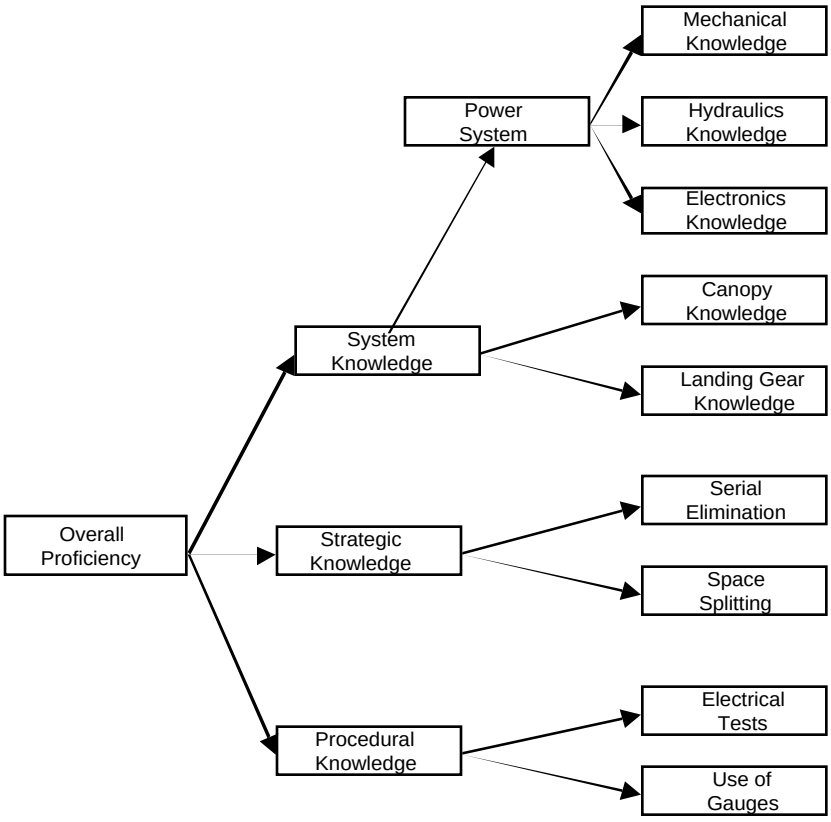


Figure 3. The student model in the HYDRIVE example is a more complex Bayes net.

that expert troubleshooting requires a conjunction of declarative, strategic, and procedural knowledge, so the student-model variables reflected key aspects of proficiency along these lines. Since the main purpose of HYDRIVE is instruction, the student model variables conform to a philosophy of instruction in the domain - in this case, troubleshooting a hierarchical physical system. The student model variables are defined at the grain-size at which instructional decisions are made - in this case, infrequent, high-level review sessions for aspects of system functionalities and troubleshooting strategies.

As discussed in the next section, we model values of observable variables as probabilistic functions of student model variables. The structure and strength of these relationships, expressed as conditional probability distributions, gives the direction and weight of evidence about student-model variables we obtain when we learn the values of observable variables. We update our belief about the person's student-model variables from what we see her do in situations that we have structured and modeled in this manner. The updated distribution of student-model variables for a given person at a given point in time can be used to trigger decisions to such as to stop testing, shift focus, offer feedback, or make a placement decision.

Evidence Models

Evidence models address Messick's second question, "What behaviors or performances should reveal [the targeted] constructs," and the natural follow-up question, "What is the connection between those behaviors and the student model variables?"

There are actually two parts of the evidence model. First, *Evidence Rules* extract the salient features of whatever the student has produced in the task situation, or the *Work Product*, and ascertain values of *Observable Variables* (Figure 4). This might be simple or complex, it might be done automatically or through human judgment.

Example 1 (continued). This is an evidence rule in the GRE P&P test:

> IF the area on the mark-sense answer sheet corresponding to the correct answer reflects more light by 10% than each area corresponding to the distractors,
>
> THEN the item response IS correct,
>
> ELSE the item response IS NOT correct.

Example 2 (continued). Here's the form evidence rules take in HYDRIVE. The work product is the list of actions a student takes in the course of working through the problem. As a student is working through a problem, short sequences of actions are grouped into clusters that each provide an item of information about the state of the system. These clusters are then evaluated in terms of their effect on the problem space.

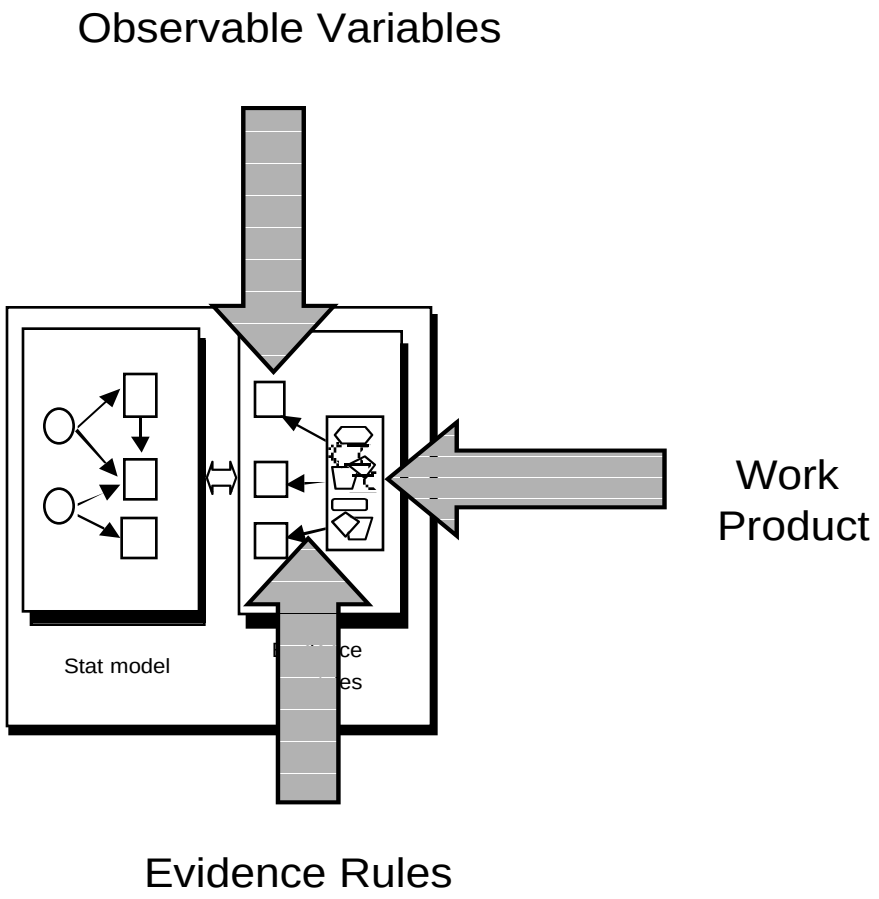


Figure 4. The *Evidence Rules* in the evidence model extract the salient features of the work product a student produces, and summarize them in terms of values for observable variables.

This rule sets the value for an observable variable that says whether or not he has taken actions consistent with space-splitting the power path (an expert strategy):

> IF an active path which includes the failure has not been created and the student creates an active path which does not include the failure and the edges removed from the problem area are of one power class,
>
> THEN the student strategy IS splitting the power path,
>
> ELSE the student strategy IS NOT splitting the power path.

A given work product may give rise to several observable variables. A single essay may be evaluated in terms of multiple aspects of language use, for

example, or a science investigation may require several interdependent steps that each contribute to a composite work product. The values of a set of observable variables in such cases is a vector-valued description of the state of the performance.

The statistical component expresses how the observable variables depend, in probability, on student model variables (Figure 5). At the point in time at which a student produces a work product, the student is posited to be in a state characterized by some unknown values of student-model variables. This state gives rise to probabilities for states of observable variables, which can be observed and as such constitute evidence about the student's state (see Haertel & Wiley, 1993, on the importance of the distinction between states of knowledge and states of performance). In Portal, we also model these relationships as Bayes-net fragments. They can be attached to the student model Bayes-net fragment to absorb the evidence. The directed edges from the student model variables to the observable variables represent conditional probability distributions.

Example 1 (continued). Figure 6 shows the statistical portion of the evidence model used in the GRE CAT, an item response theory (IRT) model. On the left is a Bayesian inference network for updating the probability distribution of the student's proficiency parameter in accordance with observing her response to a particular Item *j*. On the right is a library of all items that could be given, along with the structures necessary to dock any one with the student model in order to incorporate the evidence its response contributes. In particular, previously estimated item parameters, which define the conditional probability distribution of item responses, are available. The information stored along with these fragments also informs how to select the next item so the next response will be optimally informative while retaining the balance of kinds of items that are presented and the aspects of skill that are tapped.

Example 2 (continued). Figure 7 is the statistical part of the evidence model in the HYDRIVE example. On the left is a more complex Bayes net, in which a fragment containing two observed variables is docked with the student model, connected to the student-model variables that we posit drive their response probabilities. The structure of these fragments

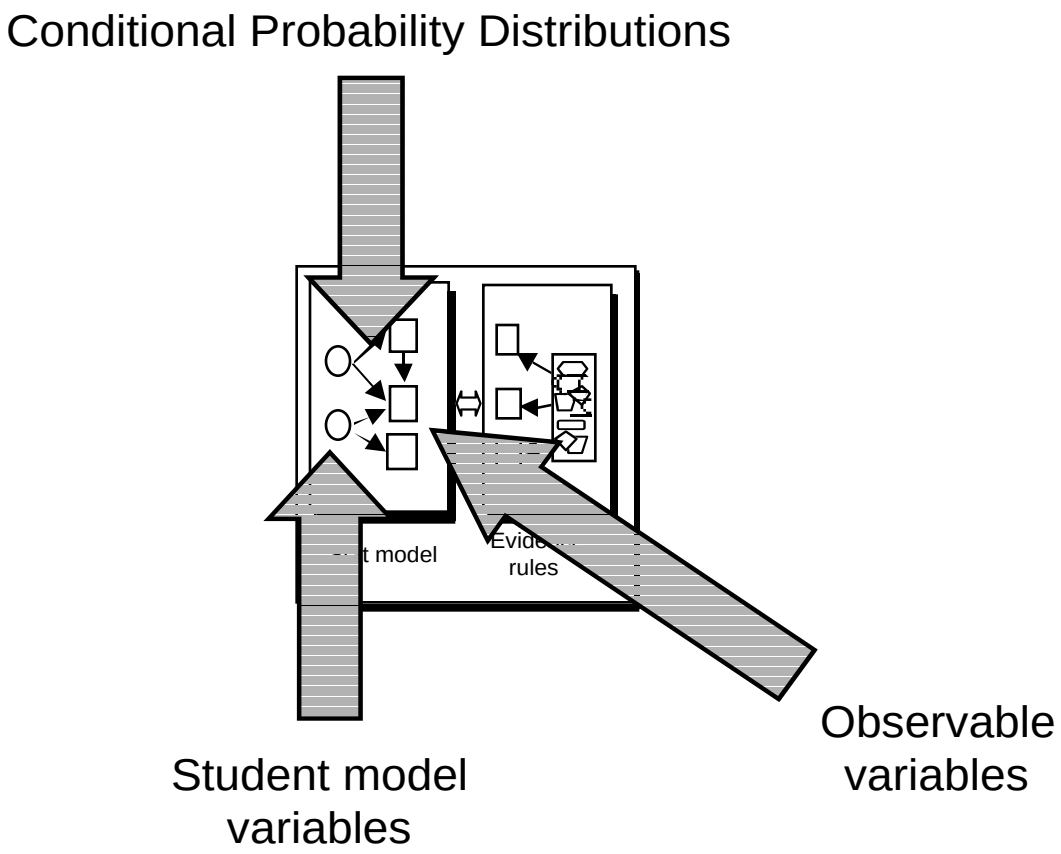
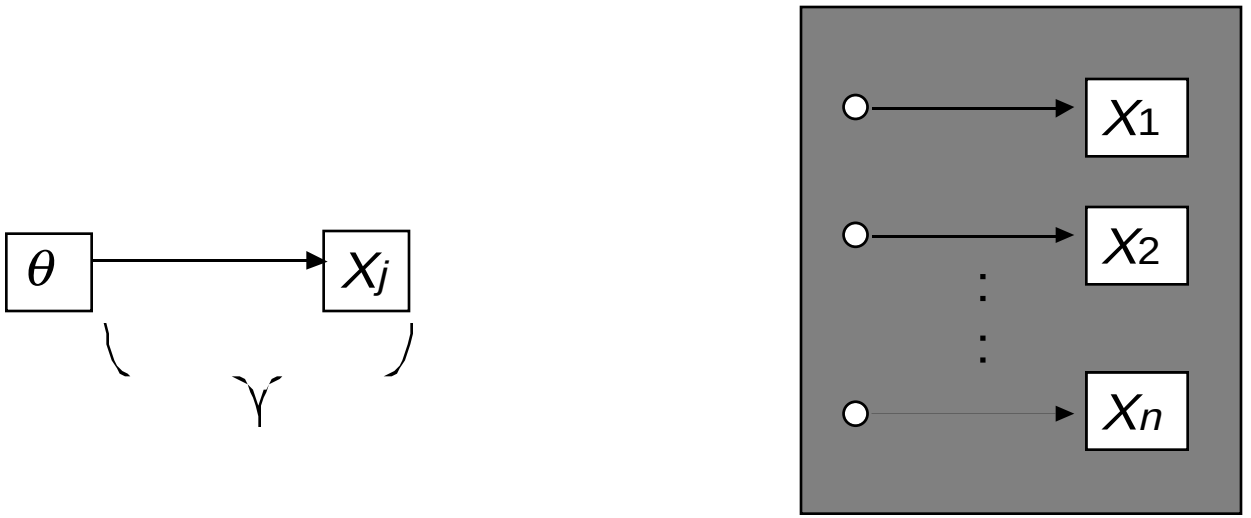


Figure 5. The *Statistical Model* in the evidence model expresses the probabilistic dependence of observable variables on student-model variables.



Sample Bayes-net fragment

(IRT model and parameters for this item)

Figure 6. The *Statistical Mo*del in the evidence model for the GRE example

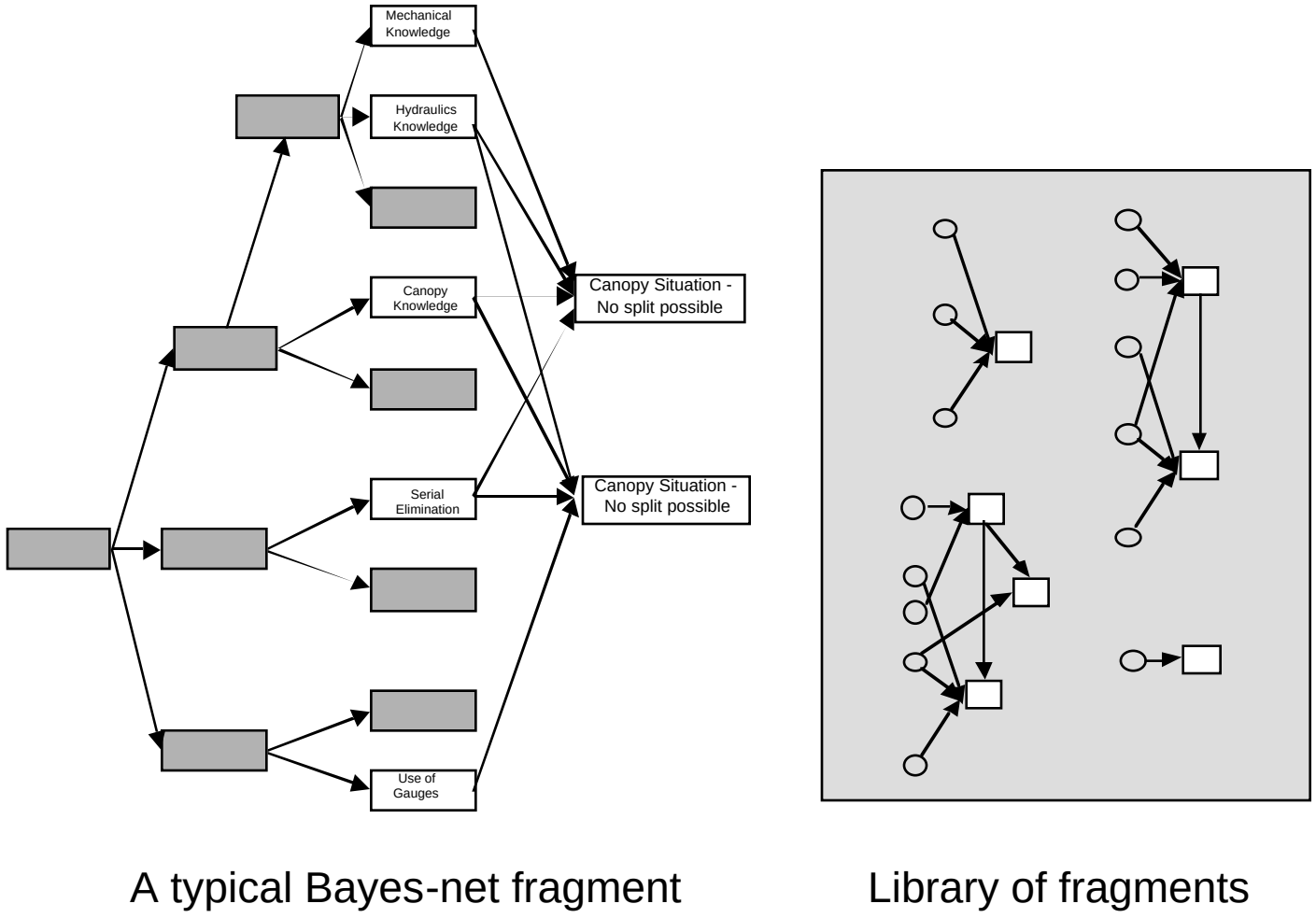


Figure 7. The *Statistical Model* in the evidence model for the HYDRIVE example

depends on our understanding of how mechanics troubleshoot this system and how the tasks are constructed. For example, a student does space-splitting consistently only if he is (a) familiar with space-splitting as a strategic technique; (b) sufficiently familiar with the system to apply the strategy; and (c) familiar with the tests and gauges he must use to carry out the strategy in a given situation. (Note that which evidence-model fragments are used in HYDRIVE depends on the situations the student works himself into, while in the GRE-CAT the evidence models are determined by the items we decide to administer.) When we observe a student's actions in a situation like this, we determine the values of observable variables using the evaluation rules, then in turn update our belief about the student model variables through the statistical model. The student model variables synthesize the information across many situations.

If, as in the GRE, the evaluation rules simply extract a single observed variable from each task that summarizes how well they've done on that task, and what we care about is a student's tendency to do well on tasks like these, then a familiar IRT model or a classical test theory model is what we use in the statistical portion of the evidence model. In more complex situations, statistical models from psychometrics can play crucial roles here as building blocks: IRT models, rating scale models, latent class models, factor models, hierarchical models, and so on. These models evolved to address certain recurring issues in reasoning about students know and can do, given what we see them do in a limited number of circumscribed situations, often captured as judgments of different people who need not agree.

As mentioned earlier, there can be multiple evidence models in a CAF. This is because different configurations of kinds of evidence and interrelationships with student-model variables might be required. Reasons for having more than one evidence model include (a) different tasks produce different kinds of work products, which need distinct sets of evidence rules; (b) different statistical models are needed to relate observed variables to student model variables; (c) different tasks and associated evidence rules produce different observable variables; (d) different subsets of student-model variables in a multivariate problem are posited to drive probabilities of different observable variables; and

(e) different values of task model variables qualitatively change the focus of evidence.

An example of an assessment that needs only one evidence model is one that uses the same single-proficiency IRT model for all items, and does not model the item parameters in terms of different task variables for different tasks. A first example of an assessment that uses a single-proficiency IRT model but needs more than one evidence model is one that has a mix of multiple-choice items and performance tasks rated on a partial-credit scale. A second example is an assessment in which different item features are used to model the parameters of different types of items.

To be used operationally, an evidence model must be compatible with both the student model and a task model. We have seen that an evidence model is compatible with a student model if all the student-model variables that determine the observed variables in the evidence model are present in the student model. A necessary condition for an evidence model to be compatible with a task model is that they share the same work-product specifications. That is, what the student produces in the task situation and what the evidence rules interrogate must be the same kind of thing.

A further condition is agreement on specified ranges of a subset of task model variables called the *scope* of the evidence model (further defined below). A given evidence model may be used with more than one task model, if the same scope, evidence rules, and structural relationships between observable variables and student-model variables are appropriate for all the task models. Similarly, more than one evidence model could be conformable with a given task model, if all the evidence models addressed the same work product and had a compatible scopes. They could apply different evaluation rules to the work product (e.g., different scoring rubrics), or they could model observable performance as a function of variables from a different student model, as appropriate to a different educational purpose (e.g., a finer grained student model when using the item for coached practice than when using it for selection).

Task Models

Task models address Messick's third question, "What tasks or situations should elicit those behaviors [that provide evidence about the targeted

knowledge]?" A task model provides a framework for describing the situations in which examinees act (Figure 8). In particular, this includes specifications for the stimulus material, conditions, and affordances, or the environment in which the student will say, do, or produce something. It includes rules for determining the values of task-model variables for particular tasks. And it also includes specifications for the *work product*, or the form in which what the student says, does, or produces will be captured. Altogether, task-model variables describe features of tasks that encompass task construction, management, and presentation. We will discuss the several roles of task model variables more fully in the next section.

Assigning specific values to task model variables, and providing materials that suit the specifications there given, produces a particular task. Assigning values or materials to only a subset of them produces a task shell. Multiple task models are possible in a given assessment. They may be employed to provide evidence in different forms, use different representational formats, or focus evidence on different aspects of proficiency. Again we postpone to the next section the role of task model variables in making these determinations.

A task thus describes particular circumstances meant to provide the examinee an opportunity to take some specific actions that will produce information about what they know or can do more generally. The task itself does not describe what we should attend to in the resulting performance or how she should evaluate what we see. This is determined by the evidence model, as described above, which needs to match on the work product it expects and on features specified in the scope of the task and evidence models. Distinct and possibly quite different evidence rules could be applied to the same work product from a given task; distinct and possibly quite different student models, befitting different purposes or conceptualizations of proficiency, could be informed by data from a given task.

> **Example 1** (continued). A task model in the GRE describes a class of test items. There is some correspondence between task models and GRE "item types" (e.g., sentence completion, passage comprehension, quantitative comparison). Different item types will generally require different task models, because different sets of variables needed to describe their distinct kinds of stimulus materials and presentation

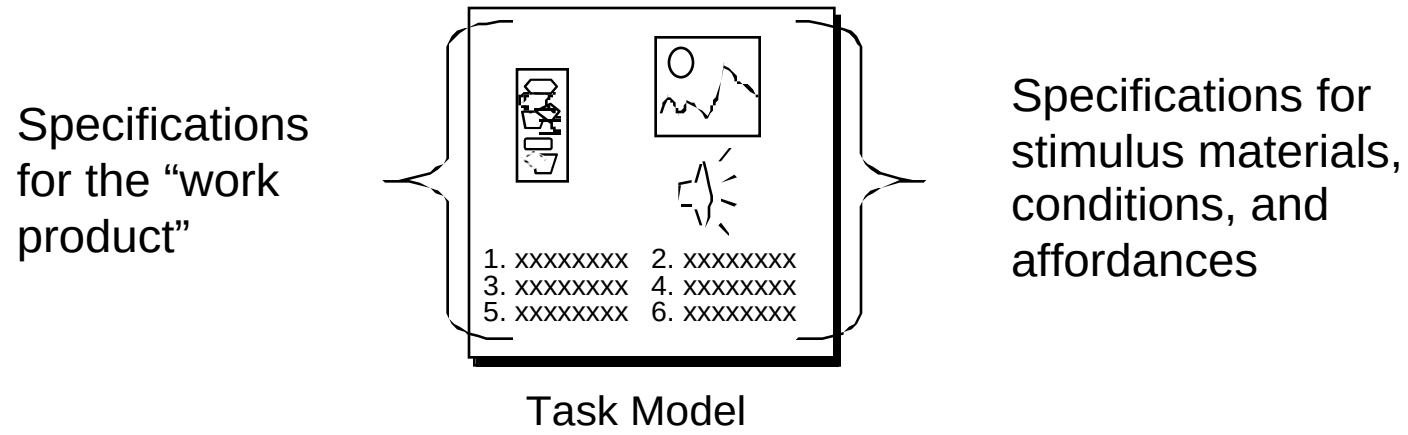


Figure 8. Elements of the Task Model

formats, and different features may be important in modeling item parameters or controlling item selection. Different task models will be required for P&P and CAT use of what is the same item from the perspective of content, because specifications for presenting and managing the item are wholly different in the two modes.

Example 2 (continued). A task model in HYDRIVE concerns the initial state of all the components in the simulation of the relevant aircraft system, video and audio clips required to present the problem and may be needed to illustrate states and actions of the aircraft during solution, and, because HYDRIVE is a coached practice system, links to an instructional module (which itself will be described in terms essentially the same as those of a task model) which can be activated by prespecified actions or states of the student model.

We will return to these task-model examples later, to say more about the variables they contain and their relationships with the other models in the assessment.

The Assembly Model

The models described above specify a domain of items an examinee might be presented, procedures for evaluating what is then observed, and machinery for updating beliefs about the values of the student model variables. *Assembly specifications* constrain the mix of tasks that constitute a given examinee's assessment. We observe neither the whole of the task domain nor an uncontrolled sample, but a composite purposefully assembled to targets for the mix of features of tasks the examinee receives. In IRT testing, optimal test assembly under multiple constraints has been a topic of much interest recently, both for fixed tests and CAT (Berger & Veerkamp, 1996). One can impose constraints that concern statistical characteristics of items, in order to increase measurement precision, or that concern non-statistical considerations such as content, format, timing, cross-item dependencies, and so on. Task selection can thus proceed with respect to constraints expressed in terms of task model variables that lie outside the statistical model proper.

The Simulator Model

The term *simulator model* refers to the capabilities that are needed to construct the environment and the situation in which the examinee will act, and to manage the interaction as may be required.

Example 1 (continued). In the GRE CAT, the simulator model is the description of requirements for the software that manages the presentation of items and captures examinees' responses. When a particular item is selected to administer to an examinee (as determined by the assembly algorithm, informed by the current state of the student model and the identification of items presented thus far), this software must render the stimulus material on the screen, provide for and respond to examinee actions such as scrolling through a reading passage, inform the examinee of time usage, and log the selected response choice. The simulator model contains the descriptions and specifications for all of these functionalities, again detailed to the extent that an external contractor could build a system that provided them.

Example 2 (continued). The HYDRIVE system contains a simulation of the hydraulic systems of the F15. It consists of objects that correspond to the mechanical, electrical, and hydraulic components of those systems, and can simulate the outcomes of troubleshooting actions in correctly functioning and variously malfunctioning states. The state of the system is updated as an examinee takes actions such as setting switches, supplying auxiliary power, replacing components, and manipulating controls. In addition, a component properly included in the simulation system tracks the implications of the student's sequence of troubleshooting actions on the so-called active path toward solution. In this way, the same action can be evaluated as *space-splitting* in one situation but *redundant* in another. Again, the simulator model for HYDRIVE contains the descriptions and specifications for these functionalities at the level appropriate to hand over to an external contractor.

ROLES OF TASK MODEL VARIABLES

In the preceding section, we described task model variables as a language for characterizing features of tasks and specifying how the interaction between the examinee and the task is managed. But what kinds of task model variables are needed, and what roles do they play? This section discusses how task model variables can play several roles in the assessment process outlined above.

Task Construction

An fundamental tenet of the evidenced-centered approach to assessment design (and of Messick's construct-centered approach as well) is that the characteristics of tasks are determined by the nature of the behaviors they must produce, to constitute evidence for the targeted aspects of proficiency. This perspective stands contrary to a task-centered approach, under which the primary emphasis is on creating tasks, with the target of inference defined only implicitly as the tendency to well on those tasks. Valuable insights inform task design under this latter approach, to be sure. But the flow of the design rationale from construct to evidence to task makes our rationale explicit from the start - easier to communicate, easier to modify, and better suited to principled generation of tasks. It is this last connection, depicted in Figure 9, we now consider.

This evidentiary perspective on assessment design also conforms nicely with a cognitive perspective on knowledge and performance. A cognitive task analysis in a given domain seeks to shed light on (a) essential features of the situations; (b) internal representations of situations; (c) the relationship between problem-solving behavior and internal representation; (d) how the problems are solved; and (e) what makes problems hard (Newell & Simon, 1972). Designing an assessment from cognitive principles, therefore, focuses on the knowledge people use to carry out valued tasks in a domain at large, and abstracts the characteristics of those tasks that provoke valued aspects of that knowledge (e.g., Embretson, 1998). Those characteristics, then, become formalized as task model variables. Irvine and his colleagues (e.g., Collis et al., 1995; Dennis et al., 1995) use the term *radical* to describe those features which drive item difficulty for theoretically relevant reasons, and *incidentals* to describe those which do not. A model for creating tasks would define variables of both types. When tasks are generated automatically, values of these variables are instantiated in a

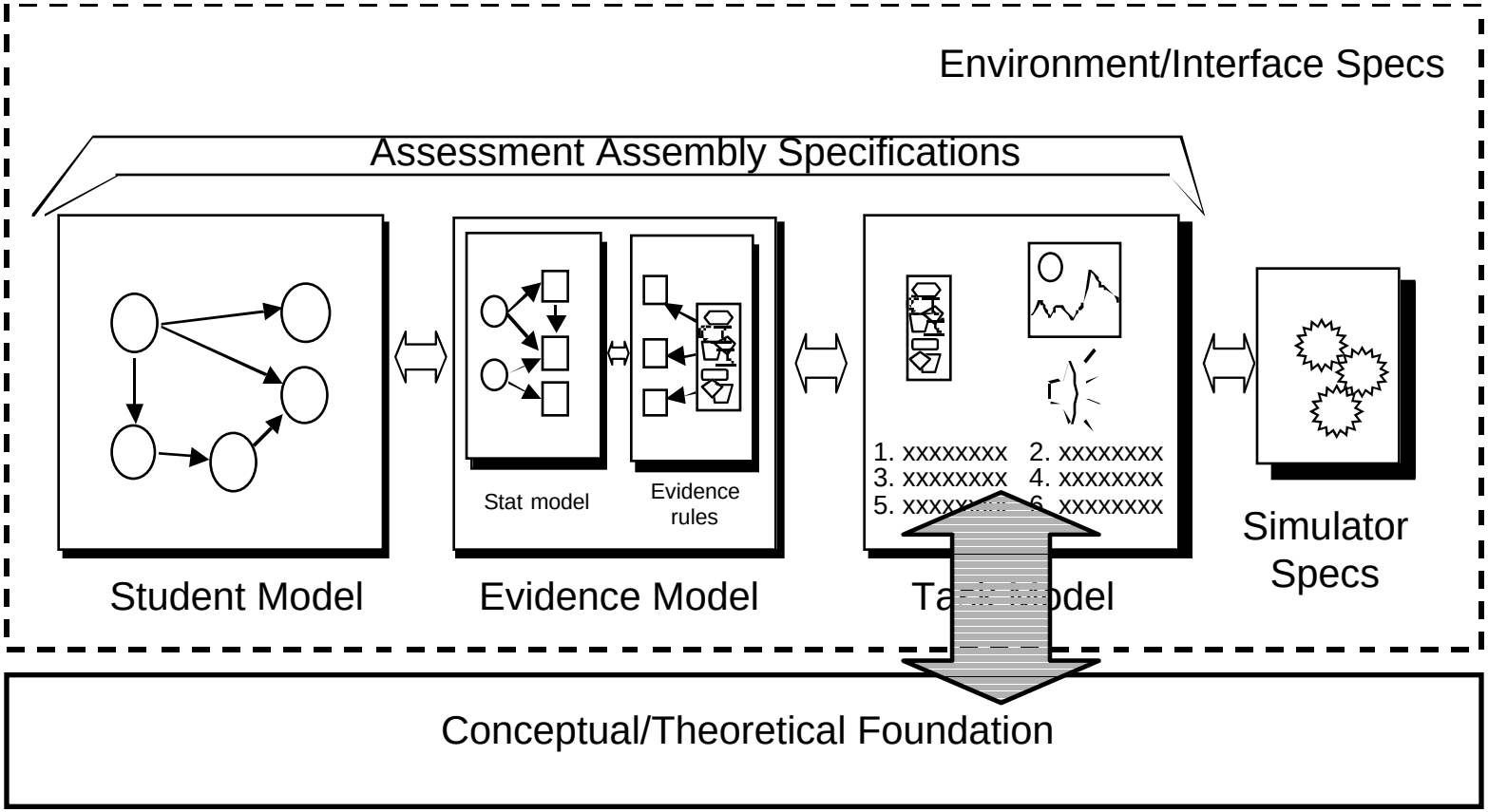


Figure 9. The role of task-model variables in *task construction* connects the operational function of creating tasks with the theoretical foundation of the assessment. What situations evoke the behavior we need to see as evidence about the proficiencies we care about?

predefined schema. The argument for the relevance of behavior in the resulting task situation is largely in place at this point, only to be verified empirical validation. Early work by Hively et al. (1968) illustrates schema-based item construction in this spirit. When tasks are created individually, values of some these task model variables are targeted when the test developer seeks stimulus material or are set after material is found and the rest of the task is written.

> **Example 1** (continued). Historically, GRE items have been essentially hand-crafted. Test developers write them to meet broad specifications and to take the form of established item types, but apply their own insights and intuitions as to sources of difficulty and regions of knowledge that will be tapped. Some researchers, such as Chaffin and Peirce (1988), undertook studies simply to make more explicit the definition and the structure of the domains of tasks that seemed to underlie the assessments. And beginning with the pioneering work of Jack Carroll (1976), other researchers launched investigations into the cognition that gives rise to performance on traditional item types, gaining psychologically-grounded insights into the features that make them difficult. These strands of research work backwards from the procedurally-defined assemblages of items that constitute a GRE to a principled explication of the domain of tasks that constitute evidence about "what the GRE measures," and a cognitive understanding of the skills and knowledge that seem to be required. Such an understanding provides a foundation for working forward, and indeed researchers are currently exploring how generative schemas, their parameters defined in terms of task-model variables, can be used to create items for the GRE (e.g., Enright & Sheehan, 1998).

> **Example 2** (continued). Given the aircraft simulator in HYDRIVE, one creates a task by specifying which components are faulty, and in which ways, among the possibilities the simulator can accommodate. The relevant task model variables thus indicate initial states for all the components.[5] When video and audio clips accompany such a task,

[5] Specifying these states was not as difficult for a task creator as it might first seem, since the normal conditions for all components were defaults. Only exception conditions had to be indicated.

additional task model variables describe technical specifications for running them and substantive characteristics relevant to the knowledge required to solve the problem (e.g., is there obvious cue to the problem? ... an irrelevant cue in the introductory clip? ... multiple cues which only together provide information about the area of the fault?).

Focusing Evidence

The proficiencies defined in any assessment have many facets, and the features of tasks can be controlled to focus the evidence on some of these facets rather than others. The *scope* of an evidence model is a list of task-model variables and associated ranges of values that must be consistent with corresponding values of those variables in a task, for that evidence model to be used to extract information from the task (Figure 10).

Example 2 (continued). An evidence model constructed to extract evidence about space-splitting usage must, perforce, be used in a task situation in which it is possible to carry out space-splitting. Its scope would include the task model variable "Space-splitting possible?" constrained to the value "Yes." Only a task model with a scope that contains "Space-splitting possible?" constrained to the value "Yes" is compatible with an evidence model that can update a "space-splitting knowledge" student-model variable.

The student model in the HYDRIVE example includes variables for knowledge about the subsystems of the hydraulics system, including the flaps and the canopy, and for knowledge about using troubleshooting strategies, including space-splitting and serial elimination. The preceding example shows how the variables on the scope of the evidence and task models serve to focus the evidentiary spotlight of tasks onto different variables within a complex student model. But the same HYDRIVE tasks could be used for in an end-of-course test with only a single overall proficiency variable in the student model. The same scope designations would still be needed, however, to ensure that an appropriate set of evidence rules was applied to extract evidence from the work product produced in response to a task. In this latter case, the variables used in the scope task serve to focus evidence-gathering on a particular region of a more broadly-construed proficiency.

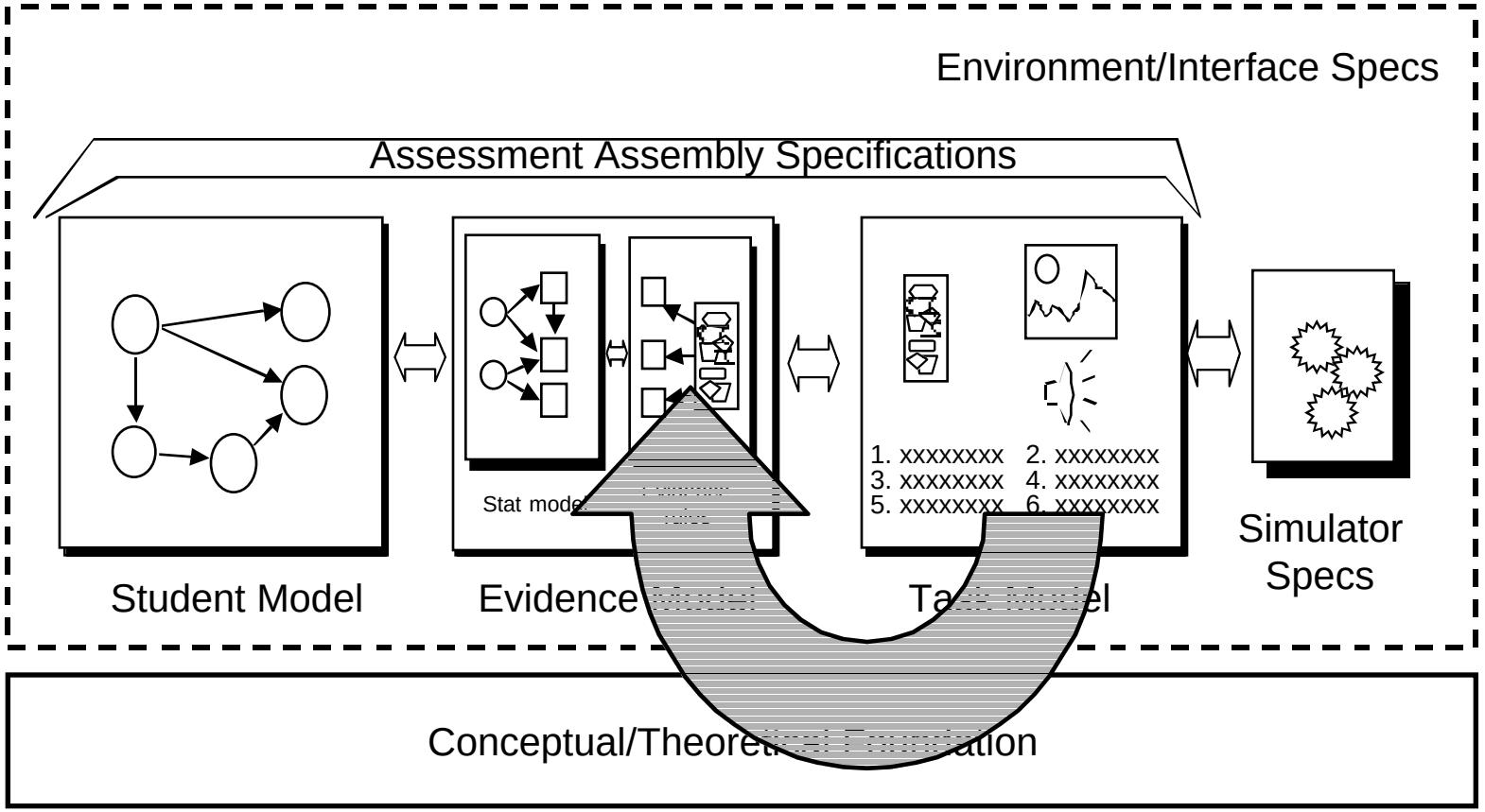


Figure 10. The role of task-model variables in *focusing evidence* connects the characteristics of task situations with the kinds of evidence that can be extracted from performance in those situations, as governed by the conceptual underpinnings of the assessment.

Assessment Assembly

Once a domain of items has been determined, test assembly specifications control the mix of tasks that constitute a given examinee's test. Constraints can be imposed with respect to statistical characteristics of tasks, such as the expected information it offers for various student model variables, and also with respect to non-measurement considerations such as content, format, timing, cross-item dependencies, and so on. In assessments that are adaptive, constraints can be dynamic as well as static. These constraints are defined in terms of task model variables, so task-model variables must be defined to characterize each task in terms of all the features needed to assembling assessments (Figure 11). The set of constraints, in conjunction with the specification of a domain of tasks, rules of evidence, and evidence-model Bayes-net fragments, constitute an operational definition of the student-model variables in an assessment - that is, they implicitly define "what the assessment measures." (A key question for the assessment designer, then, is whether this implicit functional definition of student model variables accords with the explicit conceptual definitions meant to drive the design.)

> **Example 1** (continued). In addition to information-maximizing constraints based on items' IRT parameters, the assembly specifications for the GRE CAT contain *blocking* and *overlap* constraints. Taken together, they ensure that the collection of items administered to all examinees will have similar balances of content and format, and be reasonably well modeled by the single-proficiency IRT model.
>
> Blocking constraints ensure that even though different examinees are administered different items, usually at different levels of difficulty, they nevertheless get similar mixes of content, format, modalities, skill demands, and so on. Stocking and Swanson (1993) list 41 constraints used in a prototype for the GRE CAT, including, for example, the constraint that one or two aesthetic/philosophical topics be included in the Antonym subsection. Since it is not generally possible to satisfy all constraints simultaneously, these authors employed integer programming methods to optimize item selection, with item-variable blocking constraints in addition to IRT-based information-maximizing constraints.

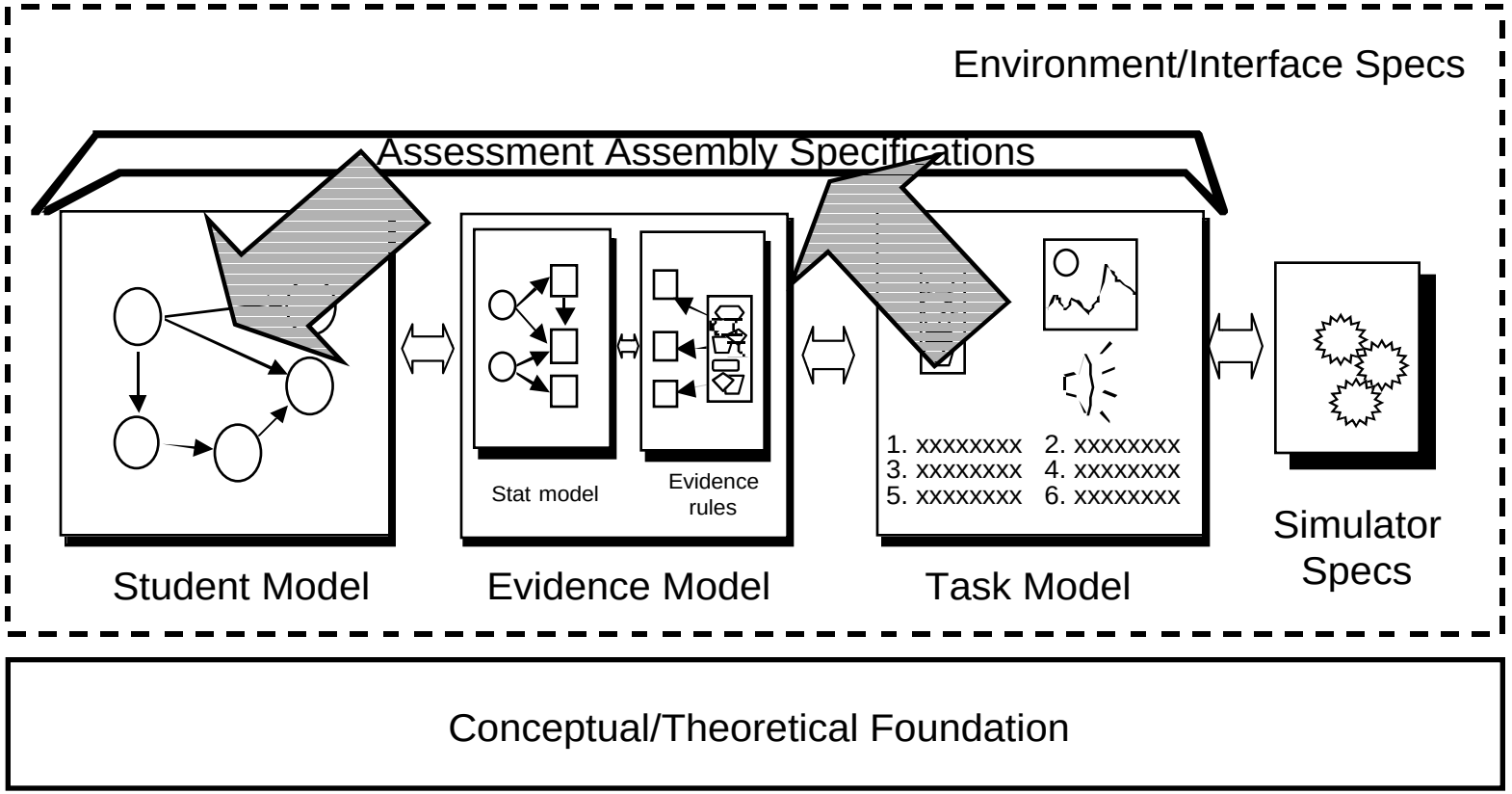


Figure 11. The role of task-model variables in *assembling assessments* connects the characteristics of task situations with the mixture of tasks which, taken together, constitute an assessment - and thereby operationally define the student-model variables.

> Overlap constraints concern the innumerable idiosyncratic features of items that cannot be exhaustively coded and catalogued. Sets of items that must not appear in the same test as one another are specified; they may share incidental features, give away answers to each other, or test the same concept. A task-model variable for GRE items, therefore, is the *enemies list*: for a particular item, this is the set of items in the same pool which cannot concomitantly appear on an examinee's test. Overlap constraints evolved through substantive lines, from the intuition that using too-similar items reduces information about examinees. Although each item is acceptable in its own right, their joint appearance causes "double counting" of evidence when a conditional-independence IRT model is used (Schum, 1994, p. 129).

IRT-CAT adapts to changing states of knowledge about the student-model variable, but the target of inference is always the same: "What is θ?" It uses information formulas and task-based blocking and overlap constraints to select items in this context. Generalizations of these kinds of item selection procedures are required for more complex models, in which different subsets of a larger set of student model variables may shift into and out of attention. Research in the psychometric literature that leads in this direction includes the work on item selection and test assembly in the context of multivariate IRT models (van der Linden, 1997; Segall, 1996) and latent class models (Macready & Dayton, 1989).

Mediating the Relationship Between Performance and Student-Model Variables

We considered above the importance of cognitively or empirically relevant features of tasks during task construction, and the role of task-model variables in structuring this process. The reason is that these features characterize which aspects of the targeted proficiencies are stressed, in which ways, and to what extent, by situations that exhibit those features. Some of these same variables can play a role in the statistical part of the evidence model for the same reason. The conditional probability distributions of the values of observable variables, given the relevant student-model variables, can be modeled as functions of these task model variables in the evidence-model Bayes-net fragments (Figure 12).

In assessments that use IRT to model conditional probabilities of observable variables given a single student-model variable, this amounts to modeling item parameters as functions of task-model variables (e.g., Fischer, 1973; Mislevy, Sheehan, & Wingersky, 1993). The practical advantage of doing this is reducing the number of pretest examinees that are needed to obtain satisfactory estimates of item parameters assembling tests and for drawing inferences about examinees. This can be done by characterizing items post hoc, but a more powerful approach is model conditional probabilities in terms of (perhaps a subset of) the same features that theory posits to be important and around which items are constructed (Bejar, 1990). Embretson (1998) illustrates how assessment design, task construction, and statistical modeling can thus be unified under a cognitive perspective. Collis et al. (1995) use the approach with computer-generated tasks, with the objective of creating items with operating characteristics sufficiently predictable to be used without any pretesting at all.

In assessments with a single student-model variable and conditionally independent observations, modeling item difficulty as a function of task model variables is closely related to the desired end of modeling conditional probabilities. Task model variables typically show similar relationships with IRT difficulty parameters and classical indices of difficulty such as percent-correct in the target population. Further, experience suggests that IRT difficulty parameters are at once easiest to model and most important in subsequent inference (Mislevy et al., op cit.).

> **Example 1** (continued). Many studies have been carried out on the features of GRE items that appear to account for their difficulty. Chalifour and Powers (1989) accounted for 62 percent of item difficulty variation and 46 percent of item biserial correlation variation among GRE analytical reasoning items with seven predictors, including the number of rules presented in a puzzle and the number of rules actually required to solve it. Scheuneman, Gerritz, and Embretson (1989) were able to account for about 65 percent of the variance in item difficulties in the GRE Psychology Achievement Test with variables built around readability, semantic content, cognitive demand, and knowledge demand.

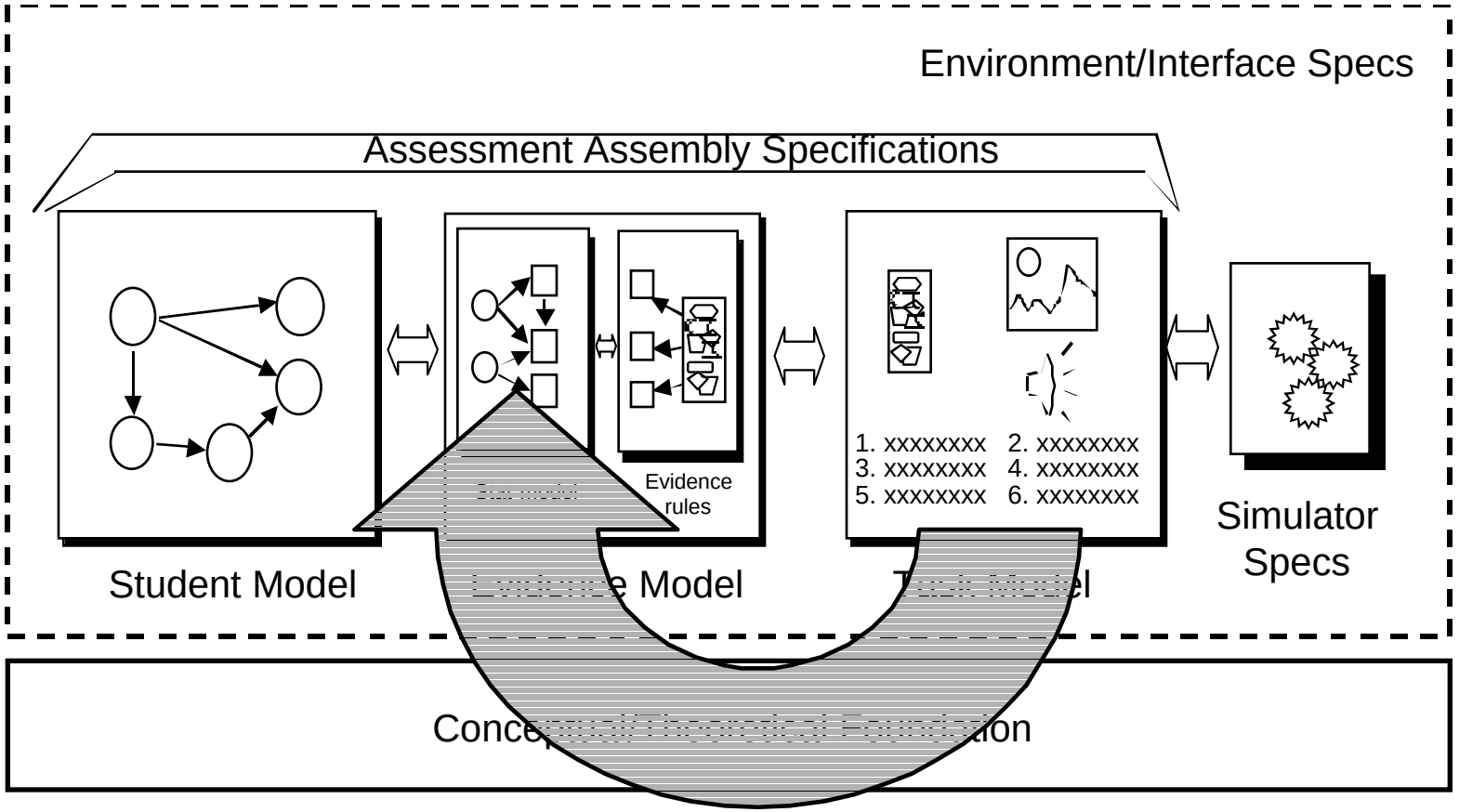


Figure 12. The role of task-model variables in *meditating the relationship between performance and student-model variables* is a technical connection between values of student-model variables and conditional probabilities of observables in task situations with specified features.

In assessments that have more than one student model variable, though, simply modeling item difficulty in terms of task model variables is not sufficient. Two tasks can be equally difficult in the sense of percents: correct in the target population, but for different reasons. If those reasons reflect differential stress on various student model variables posited to determine performance of the task, then a more complicated structure is needed to properly disambiguate the evidence about those student model variables.

> **Example 3**. The current Test of English as a Foreign Language (TOEFL) measures Reading, Listening, and Structure in three separate single-proficiency tests with conditionally independent items in each. The TOEFL 2000 development project is investigating more complex tasks that demand the use of skills across these formally separate areas. Consider a task model for a class of tasks in which an examinee must first read a passage of prose, then construct a written response to some directive based on the passage. For simplicity, suppose that only a single aspect of performance is extracted, a holistic rating of whether or not the response is both substantively appropriate and satisfactorily constructed. Finally, suppose that student model variables for Reading and Writing are posited to be drive the probabilities of this observable variable. Clearly features of both reading load and writing demand will influence the difficulty of a task in this class, in the sense of, say, proportion of satisfactory performances. But one task with a complex argument requiring a simple phase for a response, and a second task with a simple passage but demanding a formal letter for a response, could be equivalent in this regard. The conditional probabilities for a satisfactory response to the first task be low until reading skill is fairly high, but insensitive to writing skill once a low threshold is met. The conditional probabilities for a satisfactory response to the second task are a mirror image, low until a fairly high level of writing skill is present but insensitive to reading skill once a threshold is met.

Characterizing Proficiency

What does a value of a student-model variable mean? One way to answer this is by describing typical performance on various tasks in the domain from students at that level. Another role for task-model variables, then, is to link

values of student model variables to expected observable behaviors (Figure 13). This role is a corollary of the previous one, which involved modeling conditional probabilities in evidence models in terms of task model variables. The essential idea is this: Knowing the values of these task model variables for a given real or hypothetical task, one can calculate expected values of the conditional probability distributions for its observable variables in the Bayes-net fragment of a conformable evidence model. One can then calculate the corresponding probability distributions for observables for any specified values of student model variables.

Example 1 (continued). The three-parameter IRT model is used with the GRE CAT. If we know an item's parameters, we can calculate the probability of a correct student response with any given θ. We can further give meaning to a value of θ by describing the kinds of items a student at that level is likely to succeed with, and those he is not. To the extent that item features account for item parameters, then, we can describe the student's proficiency in terms of substantive task characteristics and/or cognitively relevant skills. For example, Enright, Morely, and Sheehan (in press) explained about 90 percent of the variance in item difficulty parameters in a constructed set of GRE Quantitative word problems with the factors (a) problem-type, (b) complexity, and (c) number vs. variable. A student with θ = -1 would have about a 2/3 chance of correctly answering a simple "Total cost = unit cost x units" problem presented in terms of actual numbers; a student with θ = 1 would have a 2/3 chance of success with a more complex "distance = rate x time" problem presented in terms of actual numbers; and a student with θ = 2.5 would have about a 2/3 chance with a complex cost or distance problem presented in terms of variables.

Example 2 (continued). The probability distribution of any observable variable in the HYDRIVE example depends on at least three student-model variables: one for knowledge of the subsystem involved, facility with the expert-level troubleshooting strategy that can be applied, and familiarity with the tests and procedures that apply to the situation. Behavior depends on all three, so how can the idea of behaviorally

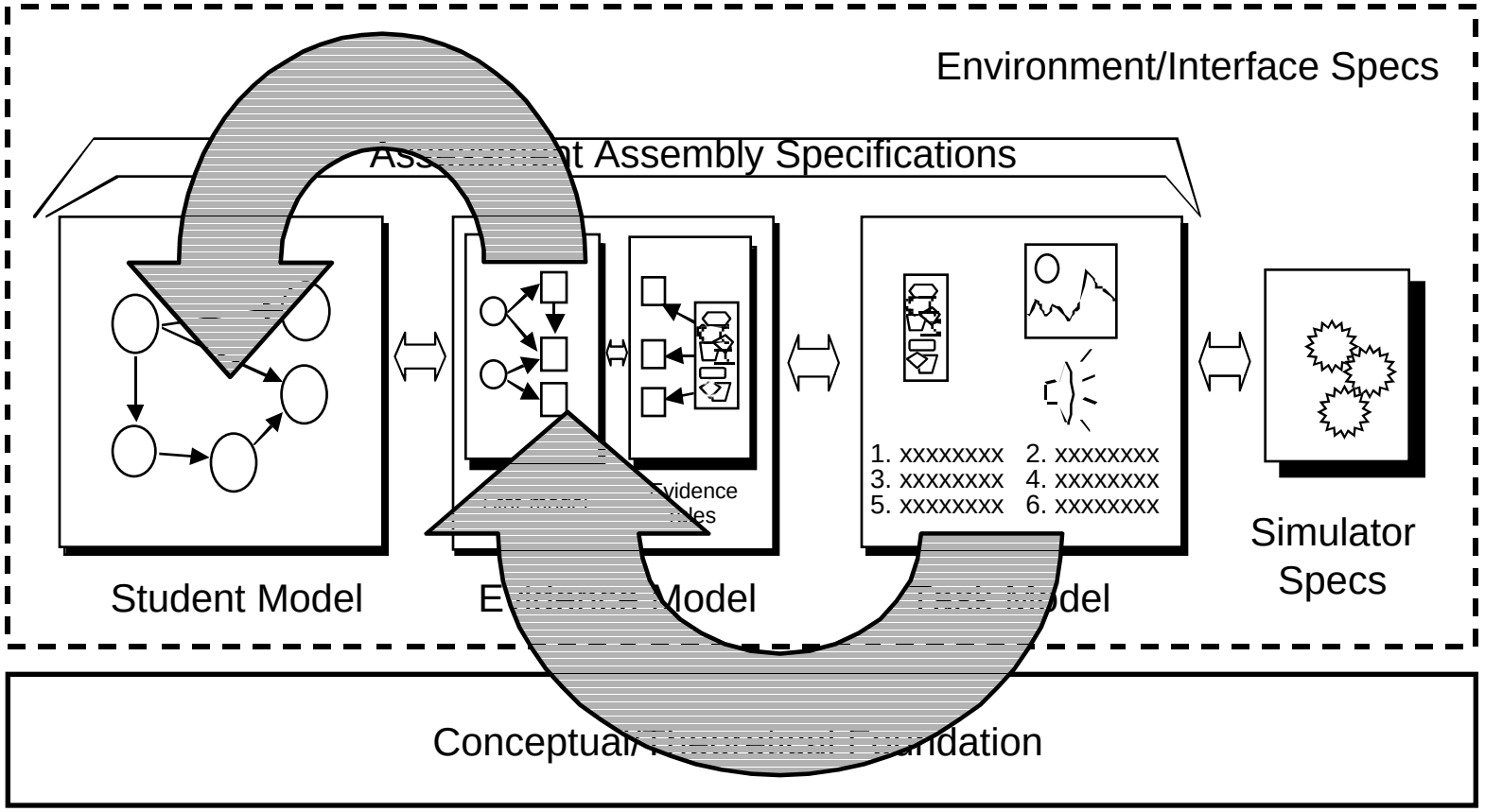


Figure 13. The role of task-model variables in *characterizing proficiency* connects values of student-model variables with expected performance in task situations with specified features, to facilitate users' interpretation of scores and exemplify the theory-driven link between scores and performances.

anchoring be applied? One approach is to identify a level of single student-model variable of particular interest and a task in which it is required, then calculate expected response probabilities for the designated level of this special student-model variable averaging over the conditional distributions of all the other student-model variables. This is a weak prediction in HYDRIVE, though, since performance depends heavily on the other student-model variables. Even if one is very good at space-splitting, he is unlikely to do it on a subsystem he is not familiar with. An alternative is to give conditional interpretations: Being high on space-splitting means a 75 percent chance of applying this expert strategy *if the examinee is familiar with the subsystem*. A second alternative is to provide descriptions of expected behavior for vectors of student-model variables. Conditional probabilities specify expectations more tightly, and the results are meaningful to the extent that the selected vectors are interpretable profiles - e.g., typical new student, or typical expert on a different aircraft.

CONCLUSION

Standard procedures for designing and carrying out assessments have worked satisfactorily for the assessments we have all become familiar with over the past half century. Their limits are sorely tested today. The field faces demands for more complex inferences about students, concerning finer-grained and interrelated aspects of knowledge and conditions under which this knowledge can be to bear. Advances in technology can provide far richer samples of performances, in increasingly realistic and interactive settings; how can we make sense of this complex data? And even with familiar assessments, cost pressures from continuous testing and social pressures for validity arguments demand more principled assessment designs and operations.

Using terms and concepts from the Portal project, we have outlined a design framework to attack these challenges. We believe that an understanding of the elements and the interrelationships that are needed for evidentiary reasoning in the assessment context provides a foundation for principled task design. We have explored the roles that variables in task models play in constructing tasks, focusing evidence, assembling assessments, characterizing proficiency, and

mediating the relationship between task performance and student proficiency. Even with such a framework, successfully designing a complex assessment remains a formidable task. Without one, though, it is almost impossible.

REFERENCES

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14,* 237-245.

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.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 27-57). Hillsdale, NJ: Erlbaum.

Chaffin, R., & Peirce, L. (1988). A taxonomy of semantic relations for the classification of GRE analogy items. *Research Report RR-87-50.* Princeton, NJ: Educational Testing Service.

Chalifour, C., & Powers, D. E. (1989). The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations. *Journal of Educational Measurement, 26,* 120-132.

Collis, J. M., Tapsfield, P. G. C., Irvine, S. H., Dann, P. L., & Wright, D. (1995). The British Army Recruit Battery goes operational: From theory to practice in computer-based testing using item-generation techniques. *International Journal of Selection and Assessment, 3,* 96-104.

Dennis, I., Collis, J., & Dann, P. (1995). *Extending the scope of item generation to tests of educational attainment.* Proceedings of the International Military Testing Association, Toronto, October, 1995.

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

Enright, M. K., Morely, M., & Sheehan, K. M. (in press). *Items by design: The impact of systematic feature variation on item statistical characteristics. GRE Research Report No. 95-15.* Princeton, NJ: Educational Testing Service.

Enright, M. K., & Sheehan, K. M. (1998). *Modeling the difficulty of quantitative reasoning items: Implications for item generation*. Paper presented at the conference "Generating items for cognitive tests: Theory and practice," co-sponsored by Educational Testing Service and the United States Air Force Laboratory, at Educational Testing Service, Princeton, NJ, November 5-6, 1998.

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

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359-384). Hillsdale, NJ: Erlbaum.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5,* 275-290.

Jensen, F. V. (1996). *An introduction to Bayesian networks*. New York: Springer-Verlag.

Macready, G. B., & Dayton, C. M. (1989, March). *The application of latent class models in adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

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

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. S. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30,* 55-78.

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

Scheuneman, J., Gerritz, K., & Embretson, S. (1989, March). *Effects of prose complexity on achievement test item difficulty*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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.

Segall, D. (1996). Multidimensional adaptive testing. *Psychometrika, 61*(2) 331-354.

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

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

van der Linden, W. J. (1997). *Multidimensional adaptive testing with a minimum error-variance criterion. Research Report 97-03.* Enschede, the Netherlands: Department of Educational Measurement and Data Analysis, University of Twente.