

PADI Technical Report 9 | July 2005



Evidence-Centered Assessment Design: Layers, Structures, and Terminology

PADI | Principled Assessment Designs for Inquiry

Robert J. Mislevy, University of Maryland

Michelle M. Riconscente, University of Maryland

Report Series Published by SRI International





SRI International
Center for Technology in Learning
333 Ravenswood Avenue
Menlo Park, CA 94025-3493
650.859.2000
<http://padi.sri.com>

PADI Technical Report Series Editors

Alexis Mitman Colker, Ph.D. *Project Consultant*
Geneva D. Haertel, Ph.D. *Co-Principal Investigator*
Robert Mislevy, Ph.D. *Co-Principal Investigator*
Laurie Fox. *Technical Writer/Editor*
Lynne Peck Theis. *Documentation Designer*

Copyright © 2005 SRI International and University of Maryland. All Rights Reserved.

Evidence-Centered Assessment Design: Layers, Structures, and Terminology

Prepared by:

Robert J. Mislevy, University of Maryland
Michelle M. Riconscente, University of Maryland

Acknowledgment

This material is based on work supported by the National Science Foundation under grant REC-0129331 (PADI Implementation Grant). We are grateful for contributions by Larry Hamel and Geneva Haertel and for comments on an earlier version by Tom Haladyna and Steve Downing.

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

CONTENTS

1.0 Introduction	1
2.0 The ECD Layers	6
2.1 Domain Analysis	7
2.2 Domain Modeling	10
2.3 Conceptual Assessment Framework	16
2.3.1 Student Model: What Are We Measuring?	18
2.3.2 Evidence Model: How Do We Measure It?	18
2.3.3 Task Model: Where Do We Measure It?	19
2.3.4 Assembly Model: How Much Do We Need to Measure It?	20
2.3.5 Sample Knowledge Representations	20
2.4 Assessment Implementation	24
2.5 Assessment Delivery	24
3.0 Conclusion	28
References	29
Appendix A: Further Reading	33
A.1 The ECD Framework	33
A.2 Applications	34
A.3 Aspects of Assessment Design and Analysis	36

FIGURES

Figure 1. Layers of Change in Buildings	4
Figure 2. Cisco Systems' Seven Layers of OSI	5
Figure 3. Graphic Representation of ECD Layers	7
Figure 4. Toulmin's (1958) Structure for Arguments	11
Figure 5. Extended Toulmin Diagram in the Context of Assessment	11
Figure 6. The Conceptual Assessment Framework (CAF)	17
Figure 7. PADI Template Objects	21
Figure 8. EDMS 738 Template Collapsed View	21
Figure 9. EDMS 738 Template in the PADI Design System	22
Figure 10. EDMS 738 Final Essay Activity in PADI Design System	23
Figure 11. EDMS 738 Length of Essay Task Model Variable in PADI Design System	23
Figure 12. The Four-Process Architecture for Assessment Delivery	25
Figure 13. Processes and Messages in the Delivery Cycle	27

TABLES

Table 1. Summary of ECD Layers	6
Table 2. Design Pattern Attributes and Corresponding Assessment Argument Components	12
Table 3. "Model Elaboration" Design Pattern in PADI Design System	13

ABSTRACT

Educational assessment is at heart an exercise in evidentiary reasoning. From a handful of things that students say, do, or make, we want to draw inferences about what they know, can do, or have accomplished more broadly. Evidence-centered assessment design (ECD) is a framework that makes explicit the structures of assessment arguments, the elements and processes through which they are instantiated, and the interrelationships among them. This presentation provides an overview of ECD, highlighting the ideas of layers in the process, structures and representations within layers, and terms and concepts that can be used to guide the design of assessments of practically all types. Examples are drawn from the Principled Assessment Designs for Inquiry (PADI) project.

1.0 Introduction

Recent decades have witnessed advances in the cognitive, psychometric, and technological tools, concepts, and theories that are germane to educational assessment. The challenge is to bring this exciting array of possibilities to bear in designing coherent assessments. This report describes a framework that facilitates communication, coherence, and efficiency in assessment design and task creation. It is the evidence-centered approach to assessment design introduced by Mislevy, Steinberg, and Almond (2003): evidence-centered design (ECD). ECD builds on developments in fields such as expert systems (Breese, Goldman, & Wellman, 1994), software design (Gamma, Helm, Johnson, & Vlissides 1994), and legal argumentation (Tillers & Schum, 1991) to make explicit, and to provide tools for, building assessment arguments that help both in designing new assessments and understanding familiar ones. This introductory section presents the principles underlying ECD. Subsequent sections describe the layers of ECD, and an Appendix provides additional resources for the theory and examples of practice that reflect the approach.

Assessment design is often identified with the nuts and bolts of authoring tasks. However, it is more fruitful to view the process as first crafting an assessment argument, then embodying it in the machinery of tasks, rubrics, scores, and the like. This approach highlights an important distinction between testing and assessment as well. Although specific tasks and collections of tasks constitute one method of gathering information relevant to an assessment, assessment is a broader term and refers to processes by which we arrive at inferences or judgments about learner proficiency based on a set of observations (American Educational Research Association, 2000). Messick (1994) sounds the keynote:

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

Messick (1994) focuses on the construct-centered approach in accordance with his purpose in writing. Salient for our purposes, however, is the chain of reasoning he identifies. Regardless of the aim of or psychological perspective assumed by a particular assessment (e.g., construct-, domain-, rubric-centered), the same chain of reasoning is central to constructing a valid assessment.

In assessment, we want to make some claim about student knowledge, skills, or abilities (KSAs),¹ and we want our claims to be valid (see Kane [in press] for a focused discussion of content-related validity evidence). Ideas and terminology from Wigmore's (1937) and

¹ Industrial psychologists use the phrase "knowledge, skills, or abilities" (KSAs) to refer to the targets of the inferences they draw. We borrow the term and apply it more broadly with the understanding that for assessments cast from different psychological perspectives and serving varied purposes, the nature of the targets of inference and the kinds of information that will inform them may vary widely in their particulars.

Toulmin's (1958) work on argumentation helps us link this goal to the concrete aspects of task development. Both used graphic representations to illustrate the fundamentals of evidentiary reasoning, the thinking that links observable but fallible data to a targeted claim by means of a warrant, a rationale or generalization that grounds the inference. In a court case, the target claim might concern whether the defendant stole a particular red car. A witness's testimony that he saw the defendant driving a red car shortly after the time of the theft constitutes evidence to support the claim, since the observation is consistent with the defendant stealing the car and driving it away—but it is not conclusive evidence, since there are alternative explanations, such as that a friend had loaned her a different red car that day. It is always necessary to establish the credentials of evidence: its relevance, its credibility, and its force (Schum, 1994).

Educational assessment reflects the same fundamental processes of evidentiary reasoning. Assessment claims concern a student's capabilities in, for example, designing science experiments, analyzing characters' motives in novels, or using conversational Spanish to buy vegetables at the market. For each claim, we need to present relevant evidence, where criteria for relevance are determined by our warrant—what we know and what we think about proficiency, and what people might say or do in particular situations that provides clues about their proficiency. Section 2.2 shows how Toulmin's and Wigmore's representational forms can be used to sketch out assessment arguments. Section 2.3 then shows how ECD moves an argument into a design for the machinery of an assessment—tasks, rubrics, statistical models, and the like—in terms of three kinds of models: Student Models, Evidence Models, and Task Models. As in law, the more complex and interrelated the collection of evidence and warrants becomes, the more helpful it is to have a framework that shows how these elements together contribute to our claim.

Another parallel to legal reasoning arises in complex cases that require a range of expertise in different disciplines. Depending on the nature of the claim, data, and warrant, it can be necessary to call upon expertise in medicine, engineering, or psychology. Communication is a crucial issue here, because within each of these disciplines a whole world of language and methods has evolved to characterize their specific types of problems. However, these languages and methods are not optimized to communicate with other "worlds." We need representations that do not constrain the sophisticated conversations and processes important for each discipline to do its work, but at the same time help us integrate key interdisciplinary conclusions into the overarching argument. A common language and a common framework are necessary to orchestrate the contributions of diverse areas of expertise. In a court case, it is the evidentiary argument that weaves together the strands of evidence and their interrelated warrants into a coherent whole (Tillers & Schum, 1991).

In assessment design, expertise from the fields of task design, instruction, psychometrics, the substantive domain of interest, and increasingly technology, are all important. Each discipline uses its own language and methods. The next section describes how the layered framework of ECD affords intradisciplinary investigations while simultaneously providing structures that facilitate communication across various kinds of expertise, each as it contributes in conjunction with the others to instantiate an assessment argument.

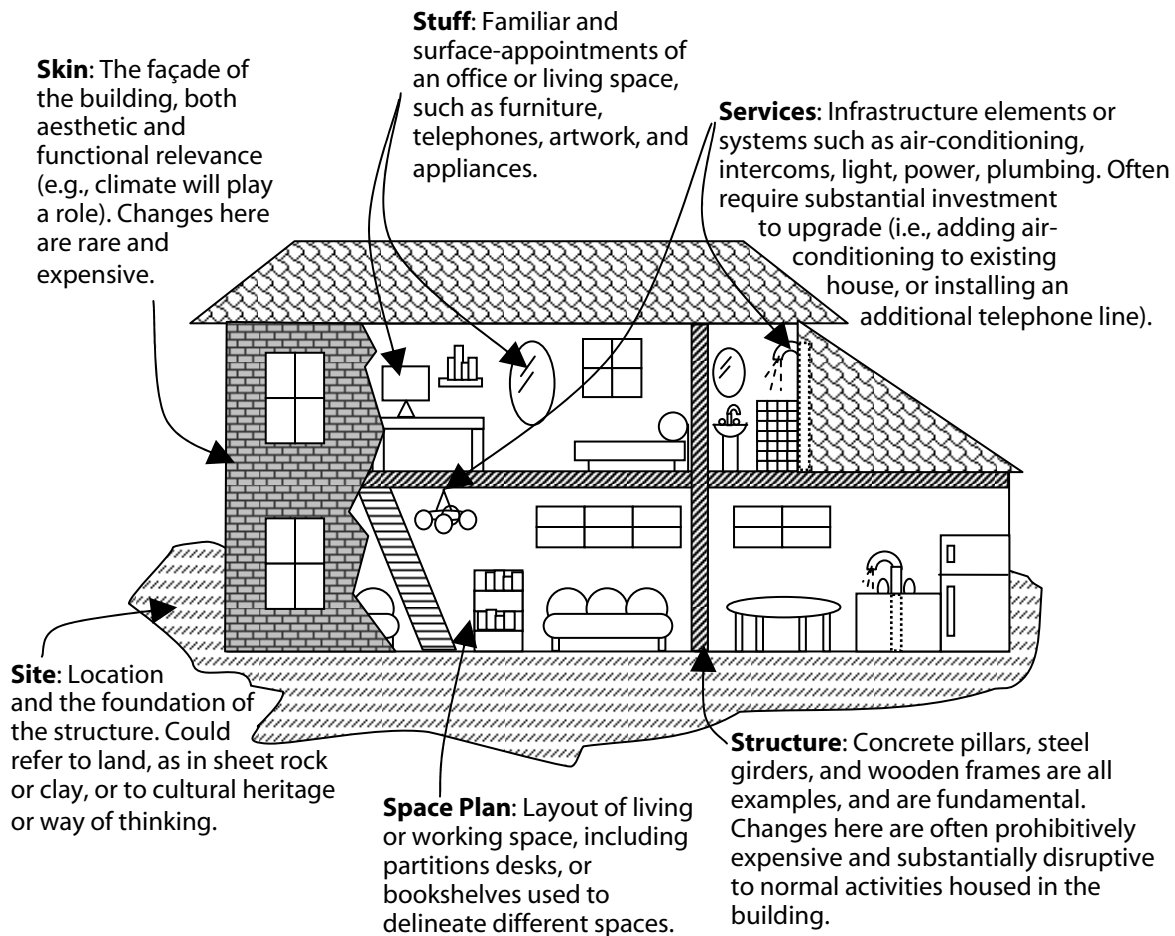
Related to the need for a common language is what we refer to as “knowledge representations” (Markman, 1998). Information, in order to be useful, must always be represented in some form. Good representations capture the important features of information in a form that people can reason with and that matches the purposes the information is to serve. For example, the city map in your car is one representation of the area and likely to be quite useful to you when lost in an unfamiliar neighborhood. The computer programming underlying Internet mapping applications (e.g., Mapquest, Yahoo Maps) doesn’t need a printed map but instead uses information presented in digital form, perhaps as a database of street attributes such as global positioning coordinates, name, and speed limit, that is tuned to route planning, transmission over the Internet, and rendering through arbitrary Web browsers. What is essentially the same information must be represented differently to be useful to different processes or people, for different purposes.

Knowledge representations are important in considerations of complex educational assessments because for various people and stages within the process, different representations of the information will be optimal: computer code for automated scoring algorithms—matrix representations of psychometric models for statisticians—and domain maps for content area specialists. The ECD framework provides domain-free schemas for organizing these knowledge representations, such as psychometric models and *task templates*, to support the construction of a solid underlying argument.

In addition to evidentiary reasoning and knowledge representations, the concept of *layers* can profitably be applied to the design and implementation of educational assessment. The compelling rationale for thinking in terms of layers is that within complex processes it is often possible to identify subsystems, whose individual components are better handled at the subsystem level (Dym, 1994; Simon, 1969). The components within these subsystems interact in particular ways, using particular knowledge representations often independent of lower-level processes elsewhere in the overall process. The subsystems are related to one another by characteristics such as timescale (as in sequential processes) for which it is possible to construct knowledge representations to support communication across subsystems as required by the overall process. Although certain processes and constraints are in place within each layer, cross-layer communication is limited and tuned to the demands of the overall goal.

Brand’s (1994) time-layered perspective on architecture provides an illustration. Drawing on the work of Frank Duffy, Brand considers buildings not as fixed objects but rather as dynamic objects wherein initial construction and subsequent change take place along different timescales, and in varying ways, by people with different roles. These layers, presented in Figure 1, serve as a heuristic for making decisions at each step in the life of a building. By employing the layers approach, activities can take place within layers that do not impact the others, but that at certain points will need to interface with adjacent layers, as when the relocation of the kitchen sink means a change of countertop, cabinet handles, and soap holders, to match the new color scheme—an interaction between Brand’s “space plan” and “stuff” layers.

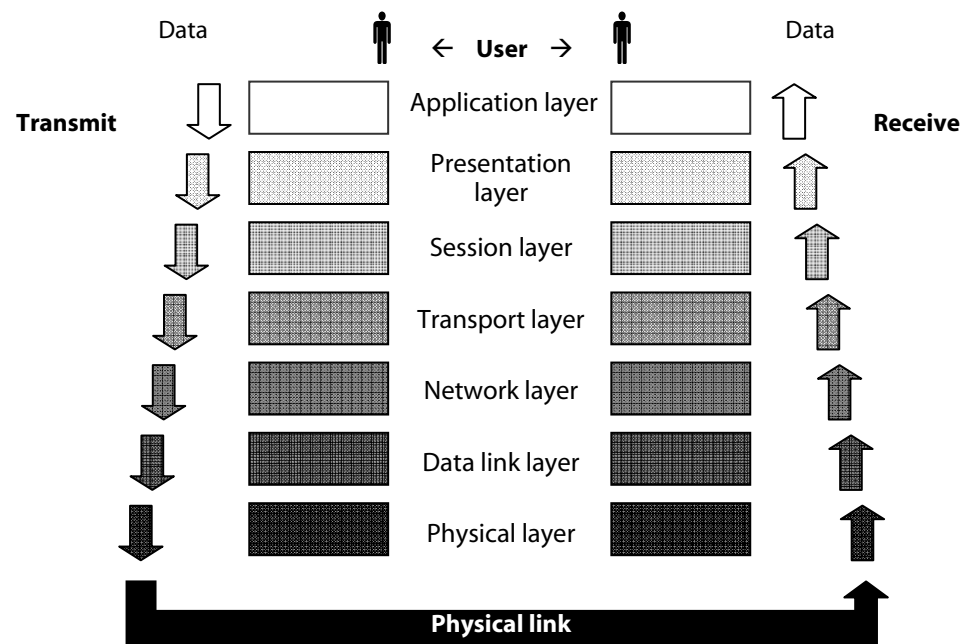
Figure 1. Layers of Change in Buildings



Use of layers is also widespread in structuring design and implementation processes in software development. A case in point is Cisco Systems' seven-layer Open System Interconnection (OSI) reference model, illustrated in Figure 2, which facilitates the transport of data from a software application on one computer to software on another computer via a network medium (Cisco, 2000):

The *Open System Interconnection (OSI) reference model* describes how information from a software application in one computer moves through a network medium to a software application in another computer. The OSI reference model is a conceptual model composed of seven layers, each specifying particular network functions.... The OSI model divides the tasks involved with moving information between networked computers into seven smaller, more manageable task groups. A task or group of tasks is then assigned to each of the seven OSI layers. Each layer is reasonably self-contained so that the tasks assigned to each layer can be implemented independently. This enables the solutions offered by one layer to be updated without adversely affecting the other layers. (p. 3)

Figure 2. Cisco Systems' Seven Layers of OSI



Reproduced from The Abdus Salam International Centre for Theoretical Physics, 1998

ECD invokes the layers metaphor in its approach to assessment. Each layer clarifies relationships within conceptual, structural, or operational levels that need to be coordinated and are either informed by, or hold implications for, other levels. Understanding the relationships within layers clarifies decision points and issues involved in making them. Although the layers might suggest a sequence in the design process, good practice typically is characterized by cycles of iteration and refinement both within and across layers.

The depictions of layers and various representations within layers discussed below are based on concepts about educational assessments discussed in Mislevy, Steinberg, and Almond (2003) and on work in the Principled Assessment Designs for Inquiry (PADI) project (Baxter & Mislevy, 2005). PADI is a National Science Foundation-sponsored project charged with developing a conceptual framework and supporting software to design science inquiry assessments. As representations, however, PADI's *design patterns* and *task templates* are applicable across content domains and educational levels.

To illustrate the application of these concepts and structures to assessment design, we use assessments from a graduate course in the foundations of assessment design, EDMS 738, as a running example. The assignments in EDMS 738 ask students to analyze aspects of actual assessments of their choice, in terms of the readings and concepts of the course. There are assignments that focus on psychological foundations of the student's example, the measurement model, the task design, and evaluation procedures. A final project requires an integrative analysis of the example incorporating all of these components.

2.0 The ECD Layers

This section walks through the ECD layers, noting the kinds of work that take place within and across layers, and offers some examples of knowledge representations in each layer. Veterans of test development are likely to find more familiar terms and concepts in the layers closest to task creation and instantiation. Therefore, a large portion of the present discussion focuses on the preceding layers in which the assessment argument is structured—the concept that guides the design choices of effective task developers but often remains in the background. Figure 3 illustrates the relationship among layers, while Table 1 summarizes the roles and key entities within the layers that are discussed in this section.

Table 1. Summary of ECD Layers

Layer	Role	Key Entities	Examples of Knowledge Representations
Domain Analysis	Gather substantive information about the domain of interest that will have direct implications for assessment, including how that information is learned and communicated.	Concepts; terminology; tools; representation forms; analyses of information use.	Various representational forms and symbol systems in a domain (e.g., algebraic notation, maps, content standards lists, syllabi).
Domain Modeling	Express assessment argument in narrative form based on information identified in Domain Analysis.	KSAs; potential work products; potential observations.	Toulmin and Wigmore diagrams; PADI design patterns.
Conceptual Assessment Framework	Express assessment argument as blueprints for tasks or items.	Student, evidence, and task models; student-model, observable, and task-model variables; rubrics; measurement models; test assembly specifications; PADI <i>templates</i> .	Algebraic and graphical representations of measurement models; PADI <i>task template</i> object model.
Assessment Implementation	Implement assessment, including presenting tasks or items and gathering and analyzing responses.	Task materials (including all materials, tools, affordances); work products; operational data for task-level and test-level scoring.	Rendering protocols for tasks; tasks as displayed; IMS/QTI representation of materials and scores; ASCII files of item parameters.
Assessment Delivery	Coordinate interactions of students and tasks; task- and test-level scoring; reporting.	Tasks as presented; work products as created; scores as evaluated.	Actual renderings of task materials in forms as used in interactions; numerical and graphical summaries for individual and group-level reports; IMS/QTI-compatible files for results.

Figure 3. Graphic Representation of ECD Layers

Domain Analysis
Domain Modeling <i>[Design Patterns]</i>
Conceptual Assessment Framework <i>[Templates and Task Specifications]</i>
Assessment Implementation
Assessment Delivery <i>[Four-Process Delivery System]</i>

2.1 Domain Analysis

The *domain analysis* layer is concerned with gathering substantive information about the domain of interest that will have implications for assessment. This includes the content, concepts, terminology, tools, and representational forms that people working in the domain use. It may include the situations in which people use declarative, procedural, strategic, and social knowledge, as they interact with the environment and other people. It may include task surveys of how often people encounter various situations and what kinds of knowledge demands are important or frequent. It may include cognitive analyses of how people use their knowledge. *Domain analysis* echoes aspects of practice analysis for credentials testing as described by Raymond and Neustel (in press). Through rich descriptions of tasks, practice analysis extracts features of tasks that are important for carrying out the responsibilities of a certain job. These task features in turn inform the kinds of student knowledge, skills, and abilities about which we will want to draw inferences as we proceed in the assessment design process.

Domain analysis also includes, at least implicitly, one or more conceptions of the nature of knowledge in the targeted domain, as Webb (in press) describes in terms of content *domain analysis* for achievement testing. How this knowledge is acquired and used, as well as how competence is defined and how it develops, will be established according to one or more psychological perspectives. Although much may be known about all of these aspects of proficiency in the targeted domain, this information usually has not been organized in terms of assessment structures. It is the substantive foundation of assessment arguments, however, and the next ECD layer—*domain modeling*—will focus on organizing the information and relationships discovered in *domain analysis* into assessment argument structures.

The psychological perspective greatly influences the overall assessment process and cannot be emphasized too strongly. The decisions regarding value and validity about knowledge and learning processes are necessarily determined according to some perspective. Why should the level of performance on a given task be useful for the assessment purpose we have in mind? Ideally, the way tasks are constructed, what students are asked to do and which aspects of their work are captured, and how their performances are summarized and reported, are all tuned to guide actions or decisions,

themselves framed in some perspective of proficiency (Embretson, 1983). The structures of ECD underscore the role of these perspectives, encouraging the test designer to make them explicit.

By way of example, imagine the domain of mathematics as seen through the lenses of the behavioral, information processing, or sociocultural perspectives (Greeno, Collins, & Resnick, 1997). In the domain of mathematics, a strict behaviorist perspective would concentrate on procedures for solving various classes of problems—possibly quite complex procedures, but ones that could be conceived of, then learned as, assemblages of stimulus-response bonds. An information processing theorist would emphasize the cognitive processes underlying acquisition of mathematics knowledge and seek to identify reasoning patterns that indicate students are on the right track as opposed to caught in common misconceptions (e.g., Siegler's [1981] balance beam tasks). A sociocultural perspective would place an emphasis on mathematics as participation in a community of practice and fluency with the forms and protocols of the domain. In each case, the situations that an assessor would design would maximize opportunities to observe students acting in ways that gave the best evidence about the kinds of inferences that were being targeted, and quite different tasks, evaluation procedures, and reports would follow.

Because the content taught, expectations of students, and modes of estimating student progress all rest on the psychological perspective assumed in instruction and assessment, it is important that this be clearly articulated throughout the assessment design process. A mismatch in psychological perspectives at different stages will result in substantially less informative assessment. The ECD approach thus suggests that assessment design entails building a coherent argument that is simultaneously consistent with the adopted psychological perspective and the claims one wants to make about examinees. Assessment design can start from a variety of points, such as claims about student proficiency (e.g., "verbal ability," as in the earlier Messick [1994] quote) or the kinds of situations in which it is important to see students doing well (e.g., Bachman & Palmer's [1996] "target language use" situations as the starting point for designing language assessment tasks), or the qualities of work at increasing levels of proficiency (e.g., Biggs & Collis's [1982] "structured outcomes of learning" taxonomy). Although the target inferences associated with different starting points will vary, all require a coherent chain of observations in order to arrive at valid claims (Kane, in press).

With this requirement in mind, we can say a bit more about the work that takes place in *domain analysis* and some organizing categories that help a designer shape the mass of information into forms that lead to assessment arguments—that is, marshalling information, patterns, structures, and relationships in the domain in ways that become important for assessment. We have noted that the psychological perspective(s) the designer assumes for the purpose of the assessment guides this process. More specifically, from the information in domain resources, we can generally identify valued work, task features, representational forms, performance outcomes, valued knowledge, knowledge structure and relationships, and knowledge-task relationships. Each of these categories has two critical features. Looking back toward the domain, they are notions that make sense to teachers, domain experts, and researchers in the domain. Looking ahead, they organize

information in ways that lead naturally to entities and structures in the next, more technical, *domain modeling* design layer.

We can identify valued work in a domain by examining real-world situations in which people engage in the behaviors and utilize the knowledge emblematic of a domain. From these situations we can ascertain the kinds of tasks appropriate for assessment, as well as discern which features of the performances themselves may be important to capture in assessment. In EDMS 738, the valued work that forms the basis of the assignments is the explication of actual and particular assessments into the conceptual framework of the ECD models, and explaining these relationships to others. Recurring and salient features of the situations in which this valued work can be observed are referred to as task features. Whereas the examinee will be in control of the performance itself, the assessment designer plays a decisive role in setting these task features in order to focus evidence, determine stress on different aspects of knowledge, and preemptively constrain alternative explanations for performance.

In any domain, information takes on a variety of representational forms depending on the nature of the content and the audience and purpose for which it is used. Learning how to use representational forms to characterize situations, solve problems, transform data, and communicate with others is central to developing proficiency in any domain. In the domain of music, for example, notation has been developed for representing compositions, with many universals and some instrument-specific features. Genetics uses Punnett squares, and mathematics uses symbol systems and operators. It is necessary to identify the representational forms—such as schematic, graphic, or symbolic systems—that accompany the target domain. Not only is much of the knowledge in the domain built into these representations, but they are used to present information and shape expectations for students' work within assessment tasks (Gitomer & Steinberg, 1999). In EDMS 738, when the students study measurement models, they must work with algebraic expressions, path diagrams, computer program interfaces, and, importantly, translate information back and forth among these forms. By identifying these representational forms, we make explicit the range of communication tools central to the domain and set the stage for using them in subsequent layers of the design process.

With performance outcomes we articulate the ways we have of knowing, appropriate to the domain of interest, when someone has arrived at an understanding or appropriate level of knowledge. That is, how do you know good work when you see it? What clues in what students say or do provide insights into the way they are thinking? These characteristics form the criteria that eventually will be necessary for crafting rubrics and scoring algorithms. Of course, characteristics of the knowledge or content of a domain also will be central to assessment design (Webb, in press). The kinds of knowledge and skill considered important in the domain are referred to as “valued knowledge.” Curriculum materials, textbooks, and concept maps of the domain are all examples of sources of valued knowledge. Of great current interest are content standards for a domain, such as the National Research Council's (1996) *National Science Education Standards*.

In addition, we may be able to specify structures and relationships underlying this valued knowledge in terms of how it tends to develop in individuals or in groups. Artifacts such as

curricula and knowledge maps provide insights into this category. Finally, we need to explicate knowledge-task relationships, meaning how features of situations and tasks interact with knowledge differences in individuals or groups. With this information we can then identify features of tasks that will prove useful for distinguishing differences in understanding between examinees. If we want to know if a student can choose an effective strategy to solve problems, we must present problems that might be approached in several ways. We must then observe whether the student uses cues in the problem setting to choose a strategy and, if this strategy founders, recognizes the signal to change strategies.

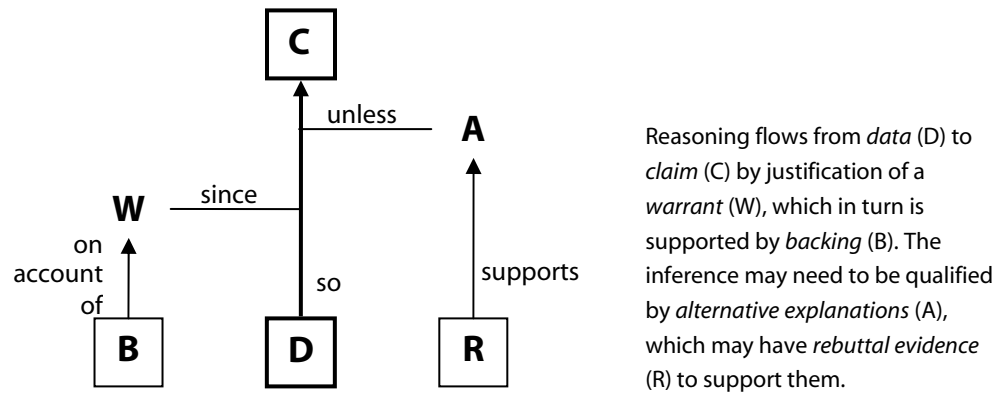
The *domain analysis* layer is furthest from the concrete tasks we ultimately seek to generate in assessment design. But the thinking along the lines sketched above underscores the importance of this layer in the overall process, building validity into assessment outcomes from the start. By making these considerations explicit, we are better able to understand existing tasks and outcomes. More importantly, we are poised to generate new tasks that will embody a grounded assessment argument.

2.2 Domain Modeling

Work in *domain analysis* identifies the elements that will be needed in an assessment. The *domain modeling* layer consists of systematic structures for organizing the content identified in *domain analysis* in terms of an assessment argument. Technical details—the nuts and bolts of particular statistical models, rubrics, or task materials—are not the focus of this layer. Rather, this layer articulates the argument that connects observations of students' actions in various situations to inferences about what they know or can do. The assessment argument takes a narrative form here: coherent descriptions of proficiencies of interest, ways of getting observations that evidence those proficiencies, and ways of arranging situations in which students can provide evidence of their proficiencies. Whereas content and instructional experts contribute the foundation of *domain analysis*, the assessment designer plays a more prominent role in *domain modeling*. Here the designer collaborates with domain experts to organize information about the domain and about the purpose of the assessment into terms and structures that form assessment arguments.

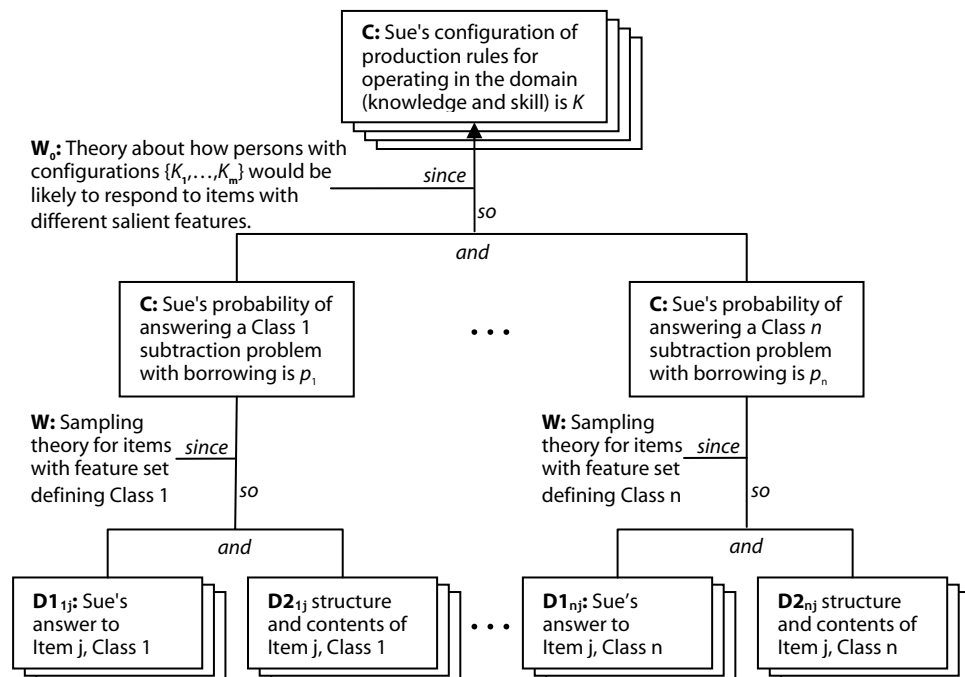
The concern of ECD at this layer is to fill in an assessment argument schema through which we can view content from any domain. Toulmin's (1958) general structure for arguments, in terms of claims, data, and warrants, provides a starting point, and Figure 4 shows this basic structure. Adapting these components to assessment design, the claim refers to the target of the assessment, such as level of proficiency in scientific problem-solving, or ability to use language appropriately in varying contexts. We provide data—such as quality of responses to questions, or behaviors observed in particular situations—to support our claims, and the warrant is the logic or reasoning that explains why certain data should be considered appropriate evidence for certain claims. Wigmore (1937) shows how evidentiary arguments in even very complicated legal cases can be expressed with assemblages of these basic structures—recurring structures in the domain, or schemas with slots to be filled.

Figure 4. Toulmin's (1958) Structure for Arguments



As an illustration, Figure 5 adapts Toulmin's and Wigmore's representations to an assessment argument. Here multiple data sources, and multiple accompanying warrants, are brought to bear on a claim about student mathematical reasoning from an information processing perspective: Sue has answered a number of subtraction problems that call for a variety of operations involving whole number subtraction, borrowing, borrowing across zeros, and so on. An information-processing perspective characterizes a student in terms of which of these operations they are able to carry out and posits that they are likely to solve problems for which they have mastered the required operations. This is the warrant, and the backing comes from both classroom experience and cognitive research, such as that of Van Lehn (1990). Patterns of responses on structurally similar tasks provide clues about the classes of problems Sue does well on and which she has trouble with. These patterns in turn provide evidence for inference about which of the operations Sue has mastered and which she has not.

Figure 5. Extended Toulmin Diagram in the Context of Assessment



PADI has adapted structures called *design patterns* from architecture (Alexander, Ishikawa, & Silverstein, 1977) and software engineering (Gamma et al., 1994; Gardner, Rush, Crist, Konitzer, & Teegarden, 1998) to help organize information from *domain analysis* into the form of potential assessment arguments (Mislevy et al., 2003). An assessment *design pattern* helps domain experts and assessment designers fill in the slots of an assessment argument. Because the structure of the *design pattern* implicitly contains the structure of an assessment argument, filling in the slots simultaneously renders explicit the relationships among the pieces of information in terms of the roles they will play in the argument. Thus, we can speak of the assessment structure as provided by the *design pattern*, and the assessment substance as determined by the assessment designer (Mislevy, 2003).

Table 2 shows the attributes of a PADI *design pattern* and their connection to the assessment argument, and Table 3 is the *design pattern* used in our running EDMS 738 example, "Model Elaboration." *Design patterns* are intentionally broad and non-technical. Centered around some aspect of KSAs, a *design pattern* is meant to offer a variety of approaches that can be used to get evidence about that knowledge or skill, organized in such a way as to lead toward the more technical work of designing particular tasks. Some other *design patterns* PADI has developed for use in assessing science inquiry include "Analyzing Data Quality," "Model Revision," "Design Under Constraints," "Self-Monitoring," and "Formulating a Scientific Explanation from a Body of Evidence."

Table 2. Design Pattern Attributes and Corresponding Assessment Argument Components

Attribute	Value(s)	Assessment Argument Component
Rationale	Explain why this item is an important aspect of scientific inquiry.	Warrant (underlying)
Focal Knowledge, Skills, and Abilities	The primary knowledge/skill/abilities targeted by this design pattern.	Student Model
Additional Knowledge, Skills, and Abilities	Other knowledge/skills/abilities that may be required by this design pattern.	Student Model
Potential observations	Some possible things one could see students doing that would give evidence about the (knowledge/skills/attributes) KSAs.	Evidence Model
Potential work products	Modes, like a written product or a spoken answer, in which students might produce evidence about KSAs.	Task Model
Characteristic features	Aspects of assessment situations that are likely to evoke the desired evidence.	Task Model
Variable features	Aspects of assessment situations that can be varied in order to shift difficulty or focus.	Task Model

Table 3. “Model Elaboration” Design Pattern in PADI Design System

Attribute	Value(s)	Comments
Title	Model Elaboration	
Summary	This design pattern concerns working with mappings and extensions of given scientific models.	A central element of scientific inquiry is reasoning with models. This design pattern focuses on model elaboration, as a perspective on assessment in inquiry and problem-solving.
Rationale	Scientific models are abstracted schemas involving entities and relationships, meant to be useful across a range of particular circumstances. Correspondences can be established between them and real-world situations and other models. Students use and gain conceptual or procedural knowledge by working with an existing model.	Students' work is bound by the concept of an existing model (or models) so that their work includes an understanding of the constraints of the problem. Even though model elaboration does not involve the invention of new objects, processes, or states, it does entail sophisticated thinking and is an analogue of much scientific activity.
Focal Knowledge, Skills, and Abilities	<ul style="list-style-type: none"> ▪ Establishing correspondence between real-world situation and entities in a given model ▪ Finding links between similar models (ones that share objects, processes, or states) ▪ Linking models to create a more encompassing model ▪ Within-model conceptual insights 	This design pattern focuses on establishing correspondences among models and between models and real-world situations.
Additional Knowledge, Skills, and Abilities	<ul style="list-style-type: none"> ▪ Familiarity with task (materials, protocols, expectations) ▪ Subject-area knowledge ▪ Reasoning within the model ▪ Model revision 	According to the designer's purposes, tasks may stress or minimize demand for other KSAs, including content knowledge, familiarity with the task type, and other aspects of model-based reasoning, including reasoning within models and revising models.

Table 3. “Model Elaboration” Design Pattern in PADI Design System (Continued)

Attribute	Value(s)	Comments
Potential observations	<ul style="list-style-type: none"> ▪ Qualities of mapping the corresponding elements between a real-world situation and a scientific model. ▪ Appropriateness of catenations of models across levels (e.g., individual-level and species-level models in transmission genetics) ▪ Correctness and/or completeness of explanation of modifications, in terms of data/model anomalies ▪ Identification of ways that a model does not match a situation (e.g., simplifying assumptions), and characterizations of the implications. 	These are examples of aspects of things that students might say, do, or construct in situations that call for model elaboration. They are meant to stimulate thinking about the observable variables the designer might choose to define for assessment tasks addressing model elaboration.
Potential rubrics		
Characteristic features	Real-world situation and one or more models appropriate to the situation, for which details of correspondence need to be fleshed out. Addresses correspondence between situation and models, and models with one another.	Any task concerning model elaboration generated in accordance with this design pattern will indicate a model or class of models the student is to work with, and real-world situations and/or other models to which correspondences are to be established.
Variable features	<ul style="list-style-type: none"> ▪ Is problem context familiar? ▪ Model provided or to be produced by student(s)? ▪ Experimental work or supporting research required? ▪ Single model or correspondence among models? ▪ How well do the models/data correspond? 	

To identify each *design pattern*, there are Title and Summary slots that summarize its purpose and basic idea. The Rationale slot articulates the underlying warrant that justifies the connection between the target inferences and the kinds of tasks and evidence that support them. Focal Knowledge, Skills, and Abilities (KSAs) come from the valued knowledge identified in *domain analysis*, and indicate the primary target of the *design pattern* (and the assessments it will be used to generate); this is the substance of the claim about students that tasks built in accordance with this *design pattern* will address. Focal as well as Additional KSAs are cast in terms of the student or examinee because our inference will concern the extent to which the student evidences them. The values of Focal and Additional KSAs are phrased as properties of a person (e.g., “ability to...,” “knowledge

of...," "skill in...," or "proficiency as needed to carry out such and such kind of work"). Additional KSAs are knowledge, skills, and abilities that might also be required in a task that addresses the Focal KSA. The task designer should consider which of these are appropriate to assume, to measure jointly, or to avoid in order to serve the purpose of the assessment. This is accomplished by design choices about variable features of tasks, as further noted below.

In the case of EDMS 738, the Focal KSA is mapping the particulars of an assessment into the form of a statistical model. Understanding the content area and language of the example assessment is an ancillary but necessary Additional KSA. The importance of the Additional KSAs becomes clear when we consider what can be inferred from a student's response to a task generated from this *design pattern*. Because a student's knowledge of the content area and language will play a role in the quality of her response, these Additional KSAs draw our attention to explanations for poor responses that are based on knowledge or skills that the task demands other than the targeted, focal, KSA—sources of "construct-irrelevant variance," to use Messick's (1989) term.

Potential Work Products are kinds of student responses or performances themselves that can hold clues about the Focal KSAs. These are things that students say, do, or make; they are thus expressed as nouns. Potential Observations concern the particular *aspects* of work products that constitute the evidence. As such, they are adjectives, describing qualities, strengths, or degrees of characteristics of realized Work Products—the evidence the work products convey about the KSAs (e.g., "number of...," "quality of...," "level of...," "kind of..."). Potential Rubrics identify the evaluation techniques that could be used or adapted to "score" work products—that is, to identify, characterize, and summarize the work products, thereby producing values of the observations. It is possible that several observations could be derived from the same work product, as in the case of an essay written about a chemical process. If the Focal KSA is cast in terms of ability to write a coherent essay, then the Potential Observations will attend to aspects of the work product such as the logical flow of topic sentences for each paragraph, not the technical quality of the explanation of the process. In contrast, an assessment in which the focal KSA is knowledge of chemical processes may not note the quality of the writing, but rather focus on the accuracy of the chemical processes described. The rubrics for arriving at these observations thus vary in accordance with the features of work that are relevant to the KSAs of interest.

With the Characteristic Features and Variable Features attributes, the assessment designer specifies aspects of the situation in which the work products are elicited. Characteristic Features imply that generally all tasks should bear these features in some form, in order to support inferences about the Focal KSA. Variable Features pertain to aspects of the task environment that the designer can choose to implement in different ways, perhaps within specified constraints. Within the constraints of the Characteristic Features, choosing different configurations of Variable Features allows a designer to provide evidence about the Focal KSA but influence the level of difficulty, the degree of confounding with other knowledge, gather more or less evidence at lesser or greater costs, and so on. One example of a Variable Feature could be the amount of scaffolding a student receives while

producing a response. Knowing the degree of scaffolding provided will be important for arriving at appropriate claims about that student's KSAs.

The *design pattern* structure does not dictate the level of generality or scope an assessment designer may choose to target in filling in the substance. Some PADI *design patterns* are special cases of more general ones. A "Problem Solving" *design pattern*, for example, could be linked to more specific *design patterns* for "Solving Well-Defined Problems" and "Solving Ill-Defined Problems." The well-defined problem can provide better evidence about carrying out problem-solving procedures, but at the cost of missing how students conceptualize problems. The ill-defined problem is better for getting evidence about conceptualization, but for students who cannot get started or who choose an inappropriate approach, there may be little evidence about how they carry out procedures. The way the designer constructs tasks, therefore, depends on which KSAs are of greatest interest. The *design pattern* helps by laying out which Characteristic Features are needed in tasks in order to learn about those KSAs. Another relationship that can exist among *design patterns* is for one *design pattern* to be comprised of components, such as a general "Model-Based Reasoning" *design pattern* linked with "Using a Given Model," "Elaborating a Model" (used in EDMS 738), and "Revising Models." *Design patterns* also can be linked with other sources of information such as references, sample tasks, and research. Because PADI focuses on middle-school science inquiry, PADI *design patterns* are linked with national science standards (e.g., National Research Council, 1996).

PADI *design patterns* also contain a slot for linking the *design pattern* to *templates*, the major design structure in the next layer of the system. As the following section describes, *templates* represent the assessment argument in terms of blueprints for the nuts and bolts of operational assessments, reflecting further design decisions that move closer to producing particular tasks.

2.3 Conceptual Assessment Framework

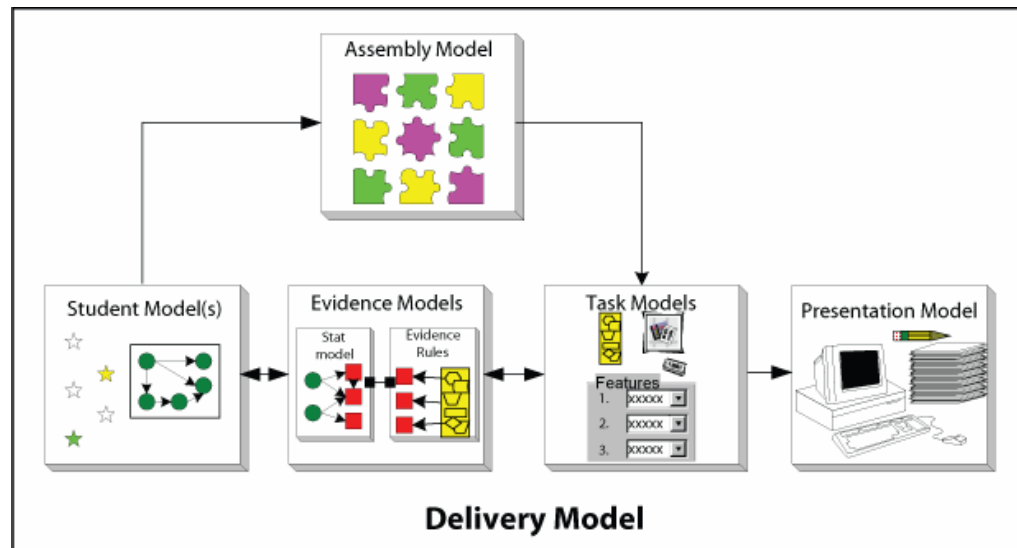
The structures in this third layer in the ECD approach to assessment design once again reflect an assessment argument, but they move away from the narrative form of *domain modeling* toward the details and the machinery of operational assessments. In the *conceptual assessment framework* (CAF) we begin to articulate the assessment argument sketched in *design patterns* in terms of the kinds of elements and processes we would need to implement an assessment that embodies that argument. The structures in the CAF are expressed as objects such as variables, task schemas, and scoring mechanisms. The substance takes the form of particular values for these variables, or for content and settings.

The CAF can be conceptualized as machinery for generating assessment blueprints by means of a structure that coordinates the substantive, statistical, and operational aspects of an assessment. In the CAF, many design decisions will be put into place to give concrete shape to the assessments we generate. These decisions include the kinds of statistical models that will be used, the materials that will characterize the student work environment, and the procedures that will be used to score students' work. When we have done the work in the CAF layer, the assessment argument will be expressed in operational terms, primed to generate a family of tasks and attendant processes that inform the target

inference about student proficiency. In addition to assessment expertise, in this layer we may also draw on technical expertise (for details of psychometric models, automated scoring, or presentation of computer-based simulations, for example) as well as on instructional expertise.

The CAF, sketched in Figure 6, is organized according to three models that correspond to the primary components of the assessment argument. These models work in concert to provide the technical detail required for implementation, such as specifications, operational requirements, statistical models, and details of rubrics. Claims, which in *design patterns* were expressed in terms of Focal and Additional KSAs, are operationalized in terms of the variables in the CAF Student Model. There can be one or several variables in a Student Model, and the Student Model can take a form as simple as an overall score across items, as complex as a multivariate item response theory or latent class model, or anything in between. What is necessary is that the Student Model Variables are the link between students' performance on tasks and the claim(s) we wish to make about student proficiency. Different values for Student Model Variables indicate different claims about students' proficiencies. A probability distribution over these variables can be used (and is used in formal probabilistic measurement models) to express what one knows about a student at a given point in time.

Figure 6. The Conceptual Assessment Framework (CAF)



The CAF Task Model comprises the components necessary to lay out the features of the environment in which the student completes the task. This is where the Characteristic and Variable Features as well as Potential Work Products from *design patterns* will be represented in terms of stimulus materials and values of the variables that describe their salient features. A variety of Potential Observations and Rubrics were identified in *design patterns*, which linked Potential Work Products to the KSAs. Each may have its own strengths and weaknesses, costs, and learning benefits. Choices among them and specific forms are now chosen to fit the purposes, resources, and context of the particular assessment that is being designed. These more specific forms are expressed in the CAF

Evidence Model. Marshalling multiple tasks into an assessment is coordinated by the Assembly Model.

2.3.1 Student Model: What Are We Measuring?

In *domain analysis* and *domain modeling*, we described targeted inference in narratives about content and student KSAs. As we have seen, it is not possible to measure these student proficiencies directly; they must instead be inferred from incomplete evidence in the form of the handful of things that students say, do, or make. The CAF lays out the statistical machinery for making inferences about student proficiencies, which will be expressed in terms of probability distributions over a single variable or set of variables.

In the simplest case, where proficiency in some defined domain of tasks is of interest, the Student Model would contain a single Student Model Variable, and students would be characterized in terms of the proportion of a domain of tasks they are likely to answer correctly. In more complex cases, where more than one proficiency is at issue, a multivariate Student Model would contain a collection of Student Model Variables and a multivariate probability distribution used to express what is known about a student's values.

The CAF contains *structures* for objects such as the Student Model, and the schemas for measurement, evaluation, and task elements discussed below. Then, for given assessments, the content or *substance* of these objects would be fleshed out: particular variables constrained either to a range or fixed value. The relationships among these objects are primed by the way the structures connect to one another.

In the EDMS 738 example, there is a single Student Model Variable. It is a continuous variable in an item response theory (IRT) model, and it is used to accumulate evidence about a student's capability to apply ECD principles to their exemplar, as evidenced by their performance on a series of assignments. A probability distribution (e.g., a maximum likelihood estimate and a standard error, or a Bayes mean estimate and a posterior standard deviation) indicates what is known about a student after evaluating their performances. This is the structure of the Student Model. The meaning of the Student Model Variable is derived from the nature of the students' performances and how they are evaluated, in particular, their reasoning about a real assessment through the lens of ECD. A simplified version of this Student Model, sufficient and appropriate for, say, classroom testing, would be to accumulate a number right or total score, and characterize its precision in terms of familiar reliability coefficients and standard errors of measurement.

2.3.2 Evidence Model: How Do We Measure It?

There are two components to the Evidence Model: the evaluation component and the Measurement Model. The first concerns the qualities of the Work Products students have produced—e.g., quality, accuracy, elegance, strategy used, and so on. The psychological perspective from which the designer views the task informs this component, since it determines the criteria for exactly which aspects of work are important and how they should be evaluated. These Observable Variables, whether quantitative or qualitative, are typically called “item scores.” Evaluation procedures are defined to say how the values of Observable Variables are determined from students' Work Products. Examples of

evaluation procedures are answer keys, scoring rubrics with examples, and automated scoring procedures in computer-based simulation tasks. Several features of a single Work Product may be important for inference, in which case evaluation procedures must produce values of multiple Observable Variables. The EDMS 738 example shows the student final essay being scored in terms of how well the students have used ECD terminology and applied ECD principles to the analyses of their chosen assessments.

Although the evaluation component tells us how to characterize the salient features of any particular performance, data like this must also be synthesized across tasks (perhaps different ones for different students) in terms of evidence for claims about what students know or can do. We need a mechanism to define and quantify the degree to which any given response reveals something about the claim we wish to make. This is the role of the Measurement Model. Each piece of data directly characterizes some aspect of a particular performance, but it also conveys some information about the targeted claim regarding what the student knows or can do. More specifically, a probability-based Measurement Model characterizes the weight and direction of evidence that Observable Variables convey about Student Model Variables. Formal psychometric models for this step include item response theory models (univariate or multivariate) and latent class models (e.g., for mastery testing). More common is the informal approximation of taking weighted or unweighted scores over items, which suffices when all items contribute relatively independent pieces of evidence about the same targeted proficiency.

2.3.3 Task Model: Where Do We Measure It?

The Task Model describes the environment in which examinees will say, do, or make something, to provide the data about what they know or can do as more broadly conceived. Decisions are made from the range of options identified in the *domain modeling* layer and expressed in *design patterns*: Potential Work Products and Characteristic and Variable Features of tasks. In the CAF layer, we specify precisely what these Work Products will be and narrow down the kinds of features that will be central or optional for grounding the targeted claims about student proficiency, under the particular constraints of the assessment situation at hand.

One decision is the form(s) the Work Product(s) should take. Will it be a multiple-choice item or an essay, for example, or a transaction list, or an illustration? What materials will be necessary as prompts for the Work Product? These include directives, manipulatives, and features of the setting such as resources available or scaffolding provided by the teacher. These features of the environment will have important implications for assessment. For example, is remembering the details of formulas a Focal KSA or not? If it is, then it is appropriate that the setting not provide this information to students so that the task will call upon their knowledge in this regard. If not, then providing open-book problems or formula sheets to students while they work is a better way to focus evidence on using formulas in practical situations. The claims about students we wish to make shape the choices of task features. Sometimes these features will be decided by students, as in the EDMS 738 example when students choose the assessment they will analyze in terms of ECD principles.

2.3.4 *Assembly Model: How Much Do We Need to Measure It?*

A single piece of evidence is rarely sufficient to sustain a claim about student knowledge. Thus, an operational assessment is likely to include a set of tasks or items. The work of determining the constellation of tasks is taken up by the Assembly Model to represent the breadth and diversity of the domain being assessed. The Assembly Model orchestrates the interrelations among the Student Models, Evidence Models, and Task Models, forming the psychometric backbone of the assessment. The Assembly Model also specifies the required accuracy for measuring each Student Model Variable. Particular forms an Assembly Model can take include a familiar test-specifications matrix, an adaptive testing algorithm (e.g., Stocking & Swanson, 1993), or a set of targets for the mix of items in terms of the values of selected Task Model Variables (e.g., the test specifications and blueprints referred to by Webb, in press).

2.3.5 *Sample Knowledge Representations*

There are various ways to implement the general description of the ECD models given above as knowledge representations that support design work in the CAF layer of the system. PADI is one of any number of systems that could be constructed as a vehicle for implementing the principles laid out by ECD. The PADI project has developed structures called *templates* (Riconscente, Mislevy, Hamel, & PADI Research Group, 2005) for doing so. Formally, a PADI *template* is the central object in the PADI Object Model, and can be represented in unified modeling language (UML; Booch, Rumbaugh, & Jacobson, 1999) or extensible markup language (XML; World-Wide Web Consortium, 1998), or in a more interactive format as Web pages in the PADI Design System (Hamel & Schank, 2005). Within an online design system, the substance of these structures is populated with definitions of Student Model Variables, Work Products, Evaluation Procedures, Task Model Variables, and the like, thereby rendering a general blueprint for a family of assessment tasks. Hierarchies of objects and their attributes are defined, which lay out specific structures and attributes for details of objects described above in more general terms as Student, Evidence, and Task Models. Figure 7 is a generic representation of the objects in a PADI *template*. Figure 8 summarizes the objects in the EDMS 738 *template* using the same graphic format. Figures 9 through 11 show parts of the actual *template* for EDMS 738 tasks from the perspective of the design system interface. In these figures, screen shots of PADI objects viewed through a Web interface are presented. Figure 9 shows part of the *template* object for the EDMS 738 example. In viewing this illustration, it is useful to recall the distinction introduced in the previous section between structure and substance. The left-most column identifies the attributes of the *template* object that define its structure; all *templates* have these attributes. The substance of each attribute is indicated in the right columns. Some of these attributes are narrative descriptions, while others are themselves objects in the PADI object model. Figures 10 and 11 present objects that are used as substance for the Activities and Task Model Variables attributes in this *template*.

Figure 7. PADI Template Objects

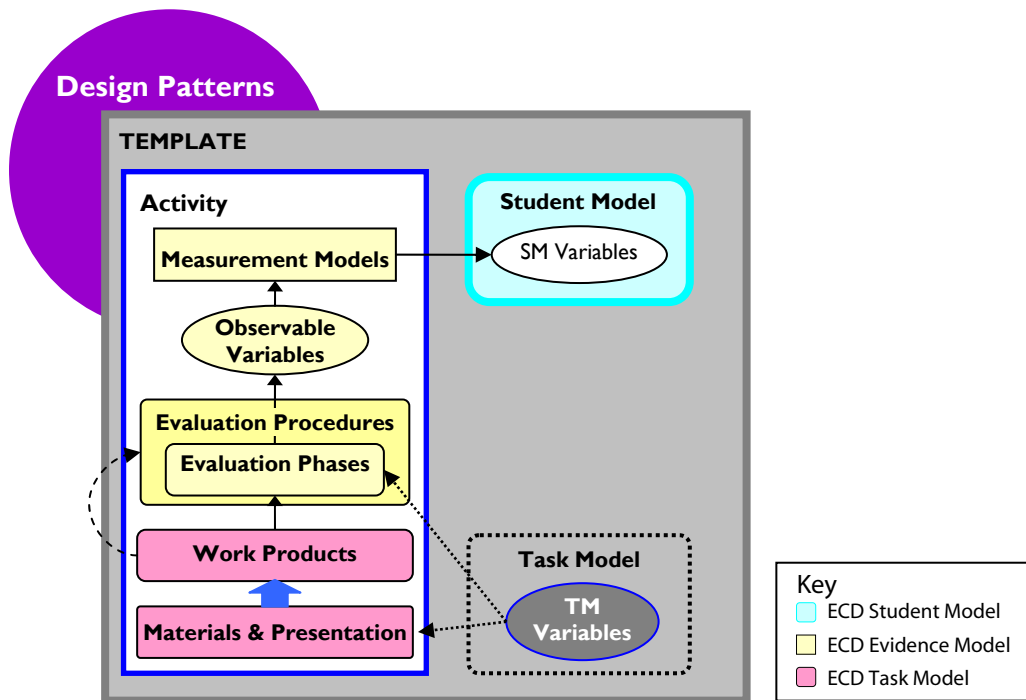
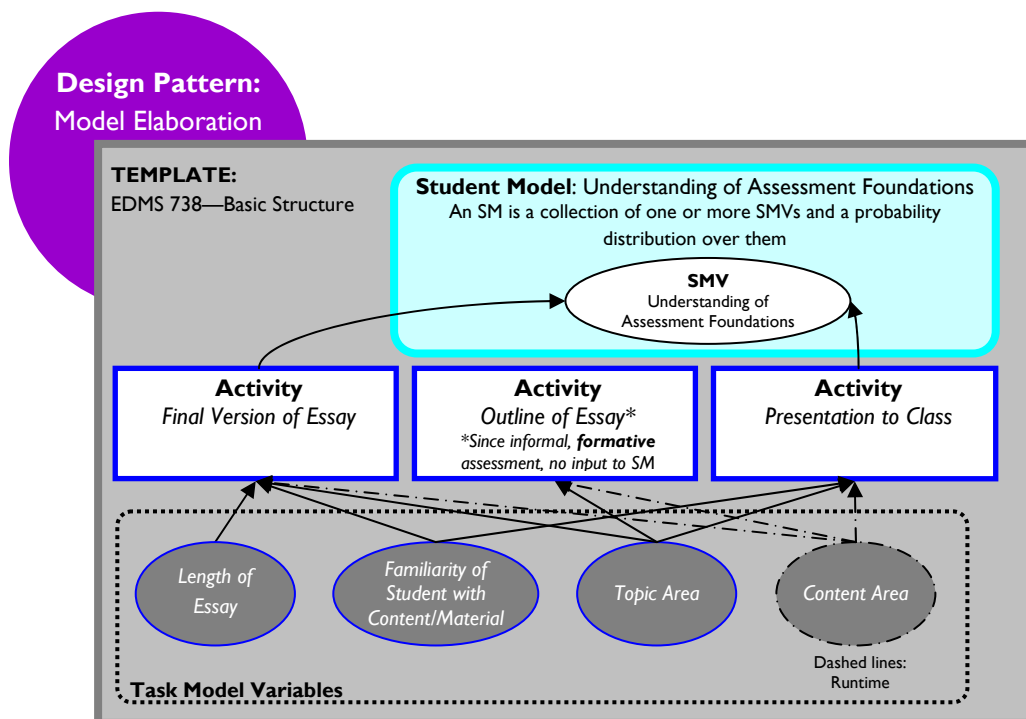



Figure 8. EDMS 738 Template Collapsed View






Design Patterns

Education Standards

Exemplars



Templates

Task Specifications

[Account Settings](#)
[Logout](#)
[Edit Mode](#)

EDMS 738 Assignments | Template 78
[[View Tree](#) | [Convert to Task Spec](#) | [Export](#)]

Title:	EDMS 738 Assignments	
Summary	Assessments for Bob Mislevy's course in Fundamentals of Assessment at U Maryland. Topics are assigned by instructor, in connection with the study, readings, and discussion of those topics through the course. Students have choice about the particular actual assessment (i.e., the 'content area') that they will analyze in their essay. The aspect(s) of assessment design, analysis, or implementation they will address in the assignment (i.e., the topic) is determined by the instructor.	required: knowledge of how to create word document on PC
Type	③ [View]	
Student Model Summary	③ one overall summary variable of proficiency	
Student Models	③ EDMS Overall Proficiency Model . Defines a univariate student model, with a continuous variable that signifies proficiency in applyin...	
Measurement Model Summary	③ univariate	
Evaluation Procedures Summary	③ generic rubrics	There are rubrics associated with the activity phases that can be used across specific topic areas.
Work Product Summary	③ Essay in MS Word format is main work product. Optional activities can produce draft outline, and in-class presentation with charts.	
Task Model Variable Summary	③	
Template-level Task Model Variables	③ topic area . topics for essay about assessment Content area . Specific domain content under consideration Amount of scaffolding . The task can guide students to think about certain concepts or can help students structure their ans... Familiarity of student with content/materials . EDMS Assignment Type . The desired kind of EDMS assignment. The list of possible responses should be the list of templates ...	
Task Model Variable Settings	③ [View]	
Materials and Presentation Requirements	③ Optional draft outline is take-home activity, can include unlimited use of materials and resources, two weeks in duration. Main activity is take-home essay, one week duration, open book. Optional class presentation is 10-minute oral presentation, with PowerPoint projection available for student's use.	
Template-level Materials and Presentation	③	
Materials and Presentation Settings	③ [View]	
Activities Summary	③ 1. (optional) review by instructor of outline by examinee 2. final draft 3. (optional) presentation to class	
Activities	③ Outline of essay . An outline of the essay is turned in to the instructor, and formative feedback is provided back to t... Presentation to class . Presentation of key points in essay to the class. Final version of essay . This is the final essay that is turned in for a grade.	
Tools for Examinee	③ computer with MS Word textbook, readings for course	
Exemplars	③	
Educational Standards	③	
Design Patterns	③ Model elaboration . This design pattern concerns working with mappings and extensions of given scientific models.	
I am a kind of	③	
These are kinds of me	③ EDMS 738 Task Spec I - Psych and Your Assessment . An assessment for Bob Mislevy's course in Fundamentals of Assessment at U. Maryland. The topic is "... EDMS 738 Task Spec II - Final Essay . This is the final assessment for Bob Mislevy's course in Fundamentals of Assessment at U. Maryland, ...	
These are parts of me	③ EDMS 738 Final Version of Essay . Assessments for Bob Mislevy's course in Fundamentals of Assessment at U Maryland. This template is ... EDMS 738 Outline of Essay . Assessments for Bob Mislevy's course in Fundamentals of Assessment at U Maryland. This template is ... EDMS 738 Presentation to Class . Assessments for Bob Mislevy's course in Fundamentals of Assessment at U Maryland. This template is ...	
Online resources	③ http://www.education.u...	Used in EDMS 738 Fall 2002, "Cognitive psychology and educational assessment"
References	③	
I am a part of	③	

Figure 10. EDMS 738 Final Essay Activity in PADI Design System

Final version of essay | Activity 82 [[View Tree](#) | [Export](#)]

Title:	Final version of essay
Summary	This is the final essay that is turned in for a grade.
Measurement Models ⓘ	Essay Grade Evidencing Proficiency . Rasch partial credit measurement fragment, modeling EDMS 738 essay grades as dependent on EDMS overa...
Evaluation Procedures ⓘ	Essay grading procedure .
Work Products ⓘ	Final Essay . Essay for final submission for grade
Materials and Presentation ⓘ	Course reading . Each course has a number of readings on course topics. One or more will be specified for a task, th... Statement of essay assignment . Textual description of assignment: Indicates form and length of required essay, aspect of assessment.. Student-selected materials describing an assessment system . In these tasks, the student chooses an assessment that is familiar to them or about which they wish ...
Presentation Logic ⓘ	Early in the course, the student will have selected a sample assessment or assessment system that they will be analyzing through the lens of the assessment design theories presented in the course. The student will gather 'student selected materials' that provide background and descriptive information about their example. When the assignment is given, the "statement of assignment" tells the student which aspect of the assessment system to analyze, and provides a list of "course readings" from which ideas are to be drawn.
Task Model Variables ⓘ	Length of essay . Long or short assignments
Design Patterns ⓘ	
Online resources ⓘ	
References ⓘ	
I am a part of ⓘ	Copy of EDMS 738 Assignments . (Template) EDMS 738 Assignments . (Template) EDMS 738 Essay with Scaf. low, topic: underpin . (Template) EDMS 738 Final Version of Essay . (Template) EDMS 738 Task Spec I - Psych and Your Assessment . (Template) EDMS 738 Task Spec II - Final Essay . (Template)

Figure 11. EDMS 738 Length of Essay Task Model Variable in PADI Design System

Length of essay | Task Model Variable 80 [[View Tree](#) | [Export](#)]

Title:	Length of essay
Summary	Long or short assignments
TMV Type ⓘ	Discrete, menu-chosen
TMV Category (possible value) ⓘ	Long: 15-25 pages Short: 2-3 pages
I am a kind of ⓘ	
These are kinds of me ⓘ	
Online resources ⓘ	
References ⓘ	
I am a part of ⓘ	EDMS 738 Task Spec I - Psych and Your Assessment . (Template) EDMS 738 Task Spec II - Final Essay . (Template) Final version of essay . (Activity)

2.4 Assessment Implementation

The next layer in the ECD assessment design scheme is *assessment implementation*. Implementation encompasses creating the assessment pieces that the CAF structures depict: authoring tasks, fitting measurement models, detailing rubrics and providing examples, programming simulations and automated scoring algorithms, and the like. Having invested expertise about the domain, assessment, instruction, and technology in a design process grounded in evidentiary reasoning, the designer is positioned to generate multiple instances of tasks from each *template*. Although these tasks may vary substantially in their surface features, having been generated from the principled assessment design process, they each embody a shared rationale and assessment argument. While most of the design decisions are finalized in this layer, some details may remain to be filled in during the subsequent layer, assessment operation. For example, mathematics tasks can be created on the fly, varying only in the values of the numbers used in identical problem structures (Bejar, 2002). In some cases these decisions can be left to the examinee, as in the EDMS 738 example where the students choose their own assessment exemplars to analyze. There, familiarity with the context and domain in an exemplar are required along with ECD principles for good analyses; letting the students choose exemplars with which they are familiar removes this additional knowledge as a source of low performance.

PADI offers support for some of the work in the implementation layer, namely specifying *templates* fully so that they are blueprints for specific tasks. These more specific structures are referred to as *task specifications*, or *task specs*. Although *templates* are capable of generating families of tasks that may vary in the range of proficiencies assessed (e.g., univariate or complex multivariate) and a host of other features such as the Observable Variables or stimulus materials, *task specs* are final plans for individual tasks. The values of some attributes will be selected from among predetermined options. Other attributes will remain unchanged, while others will have generic narrative materials tailored to their final forms.

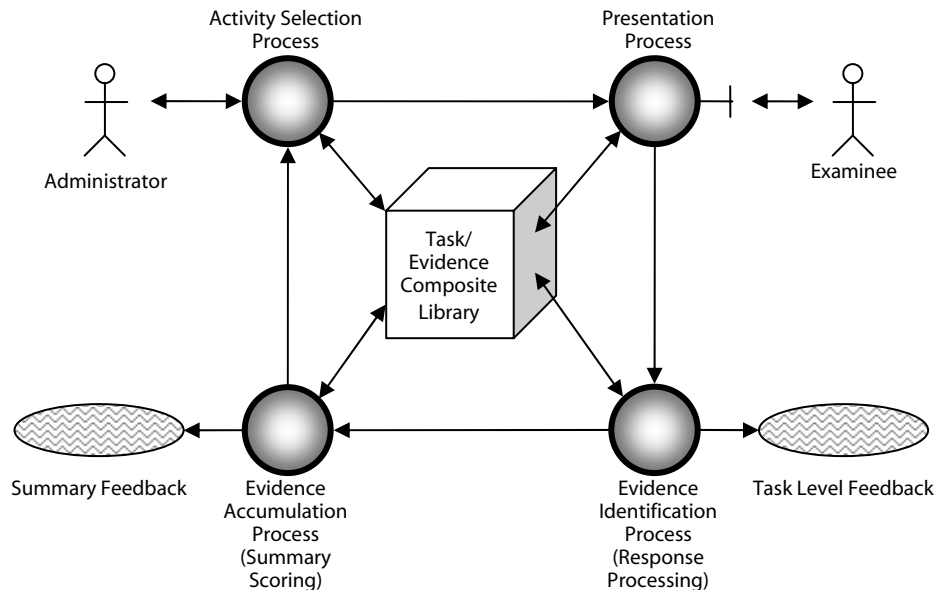
2.5 Assessment Delivery

The preceding design layers analyze a domain to determine what knowledge and skill is of interest, and how you know it when you see it; how to build an evidentiary argument from this information; how to design the elements of an assessment system that embody this argument; and how to actually build those elements. However, even the most enviable library of assessment tasks can say nothing about students in and of itself. These libraries provide only potential for learning about what students know and can do, unrealized until students begin to interact with tasks, saying and doing things that are then captured, evaluated, and synthesized into evidence about the claims at issue. Any assessment requires some processes by which items are actually selected and administered, scores are reported, and feedback is communicated to the appropriate parties.

Operational processes may differ substantially from one assessment to another, and even within a given assessment system, the processes may evolve over time as needs arise. New forms of assessment, such as computer-based simulations, require processes beyond those of familiar multiple choice and essay assessments. The international standards consortium, named Instructional Management Systems (IMS Global Learning Consortium, n.d.), has

developed Question and Test Interoperability (QTI) standards to help developers share materials and processes across assessment systems and platforms. Attention here focuses on the conceptual model of the *assessment delivery* layer that is the basis of the QTI specifications, namely the four-process architecture for *assessment delivery* shown in Figure 12 (Almond, Steinberg, & Mislevy, 2002).

Figure 12. The Four-Process Architecture for Assessment Delivery



Adapted from Mislevy, Almond, & Lukas, 2004

Assessment operation can be represented according to four principal processes. The *activity selection process* is responsible for selecting a task or other activity from the task library. In the case of our EDMS 738 example, the *activity selection process*—here, the instructor—might select the Final Essay task. This process would then send instructions about presenting the item to the presentation process, which takes care of presenting the item or task to the examinee, in accordance with materials and instructions laid out in the task model. The *presentation process* also collects responses for scoring and analysis, i.e., the Work Product(s). The Work Product may be the letter corresponding to a multiple choice option, or it may be a whole series of information, including traces of examinee navigation in an online problem-solving environment, final responses, notes, and time spent. The Work Product in the EDMS 738 example is a student’s essay, written in response to the assigned topic in the context of the examinee’s exemplar.

Work Products are passed to the *evidence identification process*, which performs item-level response processing according to the methods laid out in the Evidence Model in the CAF. This process identifies the salient outcomes of the task for the assessment purpose and expresses the outcome in terms of values of Observable Variables according to the Evaluation Procedures specified in the Evidence Model. Possibilities include the quality of writing, the accuracy of the content, or the degree to which the response reflects critical thinking. One or more outcomes can be abstracted from any given response or set of

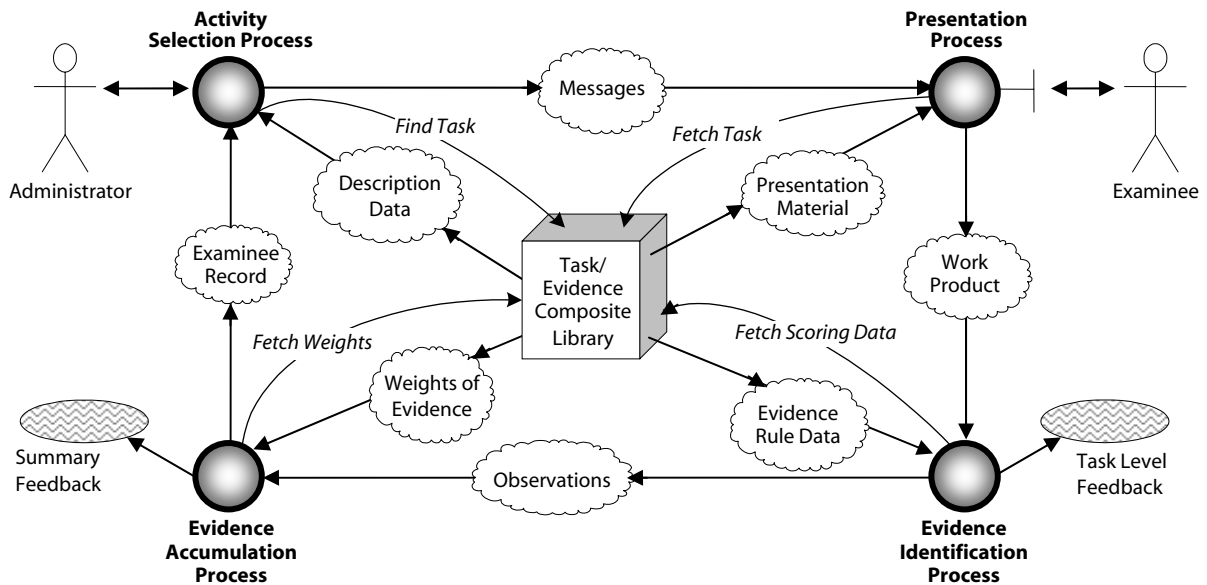
responses. Depending on the purpose of the assessment, feedback may be communicated at this point to the examinee or a teacher.

After response processing, the values of Observable Variables are sent to the *evidence accumulation process*, which is responsible for summary scoring. Here is where we amass the evidence being collected over multiple tasks in accordance with the measurement procedures specified in the CAF via the Evidence Model. This process updates the probability distributions used to express what is known about the value of a student's Student Model Variables. Summary feedback based on these results may also be provided immediately, or stored for later reporting. *Evidence accumulation* then informs the *activity selection process*, which makes a decision about the next task to administer based on criteria that may include current beliefs about examinee proficiency.

Each of these processes relies on information about how items should be presented and scored. What this information is, in abstract terms, and how it is used, was specified in the models of the CAF layer. The particulars for any given item, such as stimulus materials, item parameters, and scoring rules, were specified in the implementation layer. Now, in the operational layer, this information is stored in the *task/evidence composite library*, represented by the cube in the center of Figure 12. This library contains information about how each item should be presented, as well as parameters for how examinees will interact with the item. Conditions such as whether examinees can use calculators or spell-checkers are examples of presentation parameters. Additional information in the *task/evidence composite library* includes how responses are collected and what form they should take, as well as how to extract meaningful features from that Work Product and translate them into Observable Variables. Specifications for integrating the evidence into an accumulating student record are also contained in this library. As communication proceeds around this loop, each process will communicate directly with the *task/evidence composite library*, as well as with adjacent processes.

Figure 13 shows an expanded view of how data objects are drawn from the library and passed around the cycle. Depending on the application, a wide range of interaction patterns is possible. For example, intelligent tutoring systems, self assessment, training drills, and multiple-stage investigations would use different time-frames for responses and provide different kinds of feedback at different points in the assessment process. Further, this abstract design does not constrain the means by which processes are implemented, their locations, and their sequence and timing (e.g., the interval between evidence identification and evidence accumulation could be measured in weeks or in milliseconds).

Figure 13. Processes and Messages in the Delivery Cycle



Adapted from Mislevy, Almond, & Lukas, 2004

Now that this architecture for delivery has been defined, one can see the contribution of the IMS/QTI standards. Even though the content of the messages passed around the processes will differ substantially from one assessment system to another, and the nature and interactions of processes will vary depending on the assessment system, IMS/QTI focuses on two things that remain the same: the nature of the processes that are to be carried out, in some fashion, and the kinds of messages (data) that need to be passed from one process to another. This is sufficient to define standards for encoding these messages without restricting their content. Developers of assessment content and assessment processes can create materials and applications that are able to work together with one another because of a shared conception of the operational layer of assessment.

3.0 Conclusion

This report viewed assessment design as the development of an assessment argument, facilitated by the ECD approach. We showed how the use of layers and attention to various knowledge representations make it feasible for assessment design to coordinate work across wide ranges of expertise and technologies. To illustrate how these principles might be used in real-world assessment development, we drew on experiences and structures emerging from the PADI project (see Appendix for publications about further ECD-related theoretical considerations, as well as practical examples of ECD applications).

Current test developers have at their disposal tools such as the Toulmin structures and *design patterns* to guide their thinking about assessment design. As we sought to underscore, an essential yet often implicit and invisible property of good assessment design is a coherent evidence-based argument. Simon (1969, p. 5) refers to “imperatives” in the design of “artificial things.” Imperatives in assessment design translate into the constraints and purposes of the process. The physical nuts and bolts addressed in the *conceptual assessment framework* (CAF)—such as time limits, administration settings, and budget—may dominate considerations of constraints in the assessment design process. By engaging in the creation of *design patterns*, however, developers are supported to attend to the constraint of making a coherent assessment argument before investing resources at the CAF layer. Currently, tools at the CAF layer are still under development, with some early implementations ongoing at the Educational Testing Service and in PADI. It is our hope that in the near future off-the-shelf (or off-the-Web) supports for implementing the particulars of the processes described herein will be available. Even without software supports, however, a designer of a test at any level, in any content domain, and for any purpose, may benefit from examining test and task development from the perspective discussed here. The terminology and the knowledge representations provided in this presentation provide a useful framework for new designers and a useful supplement to experienced ones.

Initial applications of the ideas encompassed in the ECD framework may be labor intensive and time consuming. Nevertheless, the import of the ideas for improving assessment will become clear from (a) the explication of the reasoning behind assessment design decisions and (b) the identification of re-usable elements and pieces of infrastructure—conceptual as well as technical—that can be adapted for new projects. The gains may be most apparent in the development of technology-based assessment tasks, such as Web-based simulations. The same conceptual framework and design elements may prove equally valuable in making assessment arguments explicit for research projects, performance assessments, informal classroom evaluation, and tasks in large-scale, high-stakes assessments. In this way the ECD framework can serve to speed the diffusion of improved assessment practices.

References

- Abdus Salam International Centre for Theoretical Physics. (1998). *The ISO standard model for communications: OSI*. Retrieved June 5, 2005, from http://www.ictp.trieste.it/%7Eradionet/1998_school/networking_presentation/page6.html
- Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A pattern language: Towns, buildings, construction*. New York: Oxford University Press.
- Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, 1(5). Retrieved insert date, from <http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml>
- American Educational Research Association (2000). *The Standards for Educational and Psychological Testing 1999*. Washington, DC: Author.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Baxter, G., & Mislevy, R. J. (2005). *The case for an integrated design framework for assessing science inquiry* (PADI Technical Report 5). Menlo Park, CA: SRI International.
- Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Hillsdale, NJ: Erlbaum.
- Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.
- Booch, G., Rumbaugh, J., & Jacobson, I. (1999). *The Unified Modeling Language user guide*. Reading, MA: Addison-Wesley.
- Brand, S. (1994). *How buildings learn: What happens after they're built*. New York: Viking-Penguin.
- Breese, J. S., Goldman, R. P., & Wellman, M. P. (1994). Introduction to the special section on knowledge-based construction of probabilistic and decision models. *IEEE Transactions on Systems, Man, and Cybernetics*, 24, 1577-1579.
- Cisco Systems (2000). *Internetworking technology basics* (3rd ed.). Indianapolis, IN: Author.
- Dym, C. L. (1994). *Engineering design*. New York, NY: Cambridge University Press.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179-197.
- Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns*. Reading, MA: Addison-Wesley.
- Gardner, K. M., Rush, A. Crist, M. K., Konitzer, R. & Teegarden, B. (1998). *Cognitive patterns: Problem-solving frameworks for object technology*. New York: Cambridge University Press.

- Gitomer, D. H., & Steinberg, L. S. (1999). Representational issues in assessment design. In I. E. Sigel (Ed.), *Development of mental representation* (pp. 351-370). Hillsdale, NJ: Erlbaum.
- 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.
- Hamel, L., & Schank, P. (2005). *Participatory, example-based data modeling in PADI* (PADI Technical Report 4). Menlo Park, CA: SRI International.
- IMS Global Learning Consortium, Inc. (n.d.). *Question and test interoperability standards*. Retrieved June 5, 2005 from <http://www.imsglobal.org/question/index.cfm>
- Joint Committee on Standards for AERA, APA, and NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Kane, M. (in press). Content-related validity evidence in test development. In T. M. Haladyna & S. Downing (Eds.), *Handbook on test development*. Mahwah, NJ: Erlbaum.
- Markman, A. B. (1998). *Knowledge representation*. Mahwah, 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. *Educational Researcher*, 23(2), 13-23.
- Mislevy, R. J. (2003). Argument substance and argument structure. *Law, Probability, & Risk*, 2, 237-258.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Technical Report 632). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved June 6, 2005 from <http://www.cse.ucla.edu/reports/r632.pdf>
- Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., et al. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report 1). Menlo Park, CA: SRI International.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-66.
- National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.
- Raymond, M. R., & Neustel, S. (in press). Determining test content for credentialing examinations. In S. Downing & T. Haladyna (Eds.), *Handbook on test development*. Mahwah, NJ: Erlbaum.
- Riconscente, M. M., Mislevy, R. J., Hamel, L., & PADI Research Group (2005). *An introduction to task templates* (PADI Technical Report 3). Menlo Park, CA: SRI International.

- Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. New York: Wiley.
- Siegler, R. S. (1981). *Developmental sequences within and between concepts*. Monograph of the Society for Research in Child Development, Serial No. 189, 46.
- Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.
- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 277-296.
- Tillers, P., & Schum, D. A. (1991). A theory of preliminary fact investigation. *U.C. Davis Law Review*, 24, 907-966.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.
- Van Lehn, K. (1990). *Mind bugs: The origins of procedural misconceptions*. Cambridge, MA: MIT Press.
- Webb, N. (in press). Identifying content for assessing student achievement. In T. M. Haladyna & S. Downing (Eds.), *Handbook on test development*, Mahwah, NJ: Erlbaum.
- World Wide Web Consortium (1998, February 10). *Extensible Markup Language (XML) 1.0*. Retrieved from <http://www.w3c.org/TR/1998/REC-xml-19980210>.
- Wigmore, J. H. (1937). *The science of judicial proof* (3rd ed.). Boston: Little, Brown, & Co.

APPENDIX A

Further Reading

Appendix A

This report provided an overview of ECD—a start, but by no means a sufficient grounding in the subject. This appendix gives suggestions for further reading in publications that were either produced in the ECD research program or address related principles. They are presented below in three groups: publications about the ECD framework itself; applications of ideas related to ECDs; and particular aspects of assessment design and analysis from the perspective of evidentiary reasoning.

A.1 The ECD Framework

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, 1(5) 1-64. Retrieved May 1, 2005 from <http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml>
Also available as *CSE Technical Report 543*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/TECH543.pdf>

Extended discussion of the four-process delivery system architecture, including explanation of relationships between the design objects of the conceptual assessment framework and the processes and messages in an assessment delivery system.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2003). A framework for reusing assessment components. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 28-288). Tokyo: Springer.

Shorter description of the four-process delivery system, with descriptions of what the four processes do and how they interact in assessments designed to achieve different purposes.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Technical Report 632). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved June 6, 2005 from <http://www.cse.ucla.edu/reports/r632.pdf>

Brief overview of ECD, focused on the CAF and four-process delivery architecture for assessment delivery systems.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-67. Also available as *CSE Technical Report 597*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved June 6, 2005 from <http://www.cse.ucla.edu/reports/TR597.pdf>

Currently the most comprehensive overview available of evidence centered design, spanning assessment arguments, to design elements, to delivery system architecture, and the connections within and across these levels.

A.2 Applications

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global E-learning program. *The International Journal of Testing*, 4, 295-301.

Nontechnical discussion of the role and the importance of evidence-centered design in large-scale assessment programs, with a particular emphasis on those that use technology extensively. Draws lessons from Cisco Learning Institute's Cisco Networking Academy Program.

Levy, R. & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer based interactive assessment. *The International Journal of Testing*, 4, 333-369.

Focus on estimation of conditional probability models in the Bayes net psychometric model in Cisco's NetPASS prototype assessment. A fairly technical psychometric paper.

Mislevy, R. J., Almond, R. G., Dibello, L. V., Jenkins, F., Steinberg, L. S., Yan, D., & Senturk, D. (2002). *Modeling conditional probabilities in complex educational assessments* (CSE Technical Report 580). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/TR580.pdf>

Focus on estimation of conditional probability models in the Bayes net psychometric model in the Biomass prototype assessment. A fairly technical psychometric paper.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction*, 5, 253-282. Also available as *CSE Technical Report 413*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/TECH413.PDF>

Good foundational explanation of the use of Bayesian inference in complex assessments, illustrated with the HYDRIVE intelligent tutoring system for troubleshooting aircraft hydraulics.

Mislevy, R. J., Hamel, L., Fried, R., Gaffney, T., Haertel, G., Hafter, A. et al. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report 1). Menlo Park, CA: SRI International. Retrieved June 6, 2005 from http://padi.sri.com/downloads/TR1_Design_Patterns.pdf

Introduction to design patterns and illustration of the use of design patterns to support assessment of science inquiry.

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2002). Design and analysis in task-based language assessment. *Language Assessment*, 19, 477-496. Also available as *CSE Technical Report 579*. Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/TR579.pdf>

ECD perspective on designing task-based language assessments. Includes examples of Bayes nets for tasks that tap multiple aspects or knowledge and skill.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior*, 15, 335-374. Also available as *CSE Technical Report 487*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/TR487.pdf>

Design and conduct of a cognitive task analysis of expertise in dental hygiene, from the perspective of informing the construction of the models in the ECD conceptual assessment framework.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education*, 15, 363-378. Also available as *CSE Technical Report 538*. Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/RML%20TR%20538.pdf>

Argument that the way to design and analyze complex assessments, such as computer-based simulations, is from the perspective of the evidentiary argument—not from the perspective of technology. Ideas are illustrated in some detail with the DISC prototype assessment of problem-solving in dental hygiene.

Riconscente, M. M., Mislevy, R. J., Hamel, L., & PADI Research Group (2005). *An introduction to PADI task templates* (PADI Technical Report 3). Menlo Park, CA: SRI International. Retrieved June 6, 2005 from http://padi.sri.com/downloads/TR3_Templates.pdf

Describes the role of task templates in assessment design and the objects which comprise them as currently implemented in the context of the ECD-based PADI project.

Songer, N., & Wenk, A. (2003, April). *Measuring the development of complex reasoning in science*. Paper presented at the annual meeting of the American Educational Research Association, 2003, Chicago, IL.

Describes the role of ECD design principles in the BioKIDS science inquiry program.

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.

Concerns the interplay among cognitive analysis, instructional strategy, and assessment design, in the context of the HYDRIVE intelligent tutoring system for troubleshooting aircraft hydraulics.

Steinberg, L. S., Mislevy, R. J., Almond, R. G., Baird, A. B., Cahallan, C., DiBello, L. V., et al. (2003). *Introduction to the Biomass project: An illustration of evidence-centered assessment design and delivery capability* (CSE Technical Report 609). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved June 14, 2005 from <http://www.cse.ucla.edu/reports/R609.pdf>

Design rationale for a standards-based, web-delivered assessment of science inquiry, in the areas of transmission genetics and microevolution. Much discussion of working with experts and National Science Education Standards, to carry out the ECD design work and then implement a prototype assessment at the level of secondary science.

Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., & Behrens, J. T. (2004). Design rationale for a complex performance assessment. *The International Journal of Testing*, 4, 303-332.

ECD design rationale for a simulation-based assessment of troubleshooting and design of computer networks. Foundational analysis for the NetPASS on-line assessment of networking skill, by the Cisco Learning Institute, Educational Testing Service, and the University of Maryland. Includes expert-novice analysis of problem-solving.

A.3 Aspects of Assessment Design and Analysis

Almond, R. G., Herskovits, E., Mislevy, R. J., and Steinberg, L. S. (1999). Transfer of information between system and evidence models. In D. Heckerman & J. Whittaker (Eds.), *Artificial Intelligence and Statistics 99* (pp. 181-186). San Francisco: Morgan Kaufmann. Also available as *CSE Technical Report 480*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/TECH480.pdf>

Concerns the technical issue of maintaining student-model and measurement-model fragments of Bayes nets, to be assembled dynamically as is required in adaptive assessments.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, 23, 223-237. Also available as *CSE Technical Report 434*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/TECH434.PDF>

Early discussion of the kinds of variables that arise in language assessment, the roles they play in the assessment argument, and where they fit in with Bayes net modeling of performance.

Collins, A. & Ferguson, W. (1993). Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational Psychologist*, 20(1), 25-42.

Defines epistemic forms as structures which guide inquiry, and epistemic games as the strategies for filling them in.

DeMark, S. F. & Behrens, J. T. (2004). Using statistical natural language processing for understanding complex responses to free-response tasks. *The International Journal of Testing*, 4(4), 295-301.

Discussion of the use of automated evaluation procedures from natural language processing in the context of the ECD-designed NetPASS simulation-based assessment.

Gitomer, D. H., & Steinberg, L. S. (1999). Representational issues in assessment design. In I. E. Sigel (Ed.), *Development of mental representation* (pp. 351-370). Hillsdale, NJ: Erlbaum.

Discussion of the key role of representational forms in assessment. Addresses both the use of representational forms to provide information and elicit responses from examinees, and the role of assessments as representations themselves as to what is important in a domain and how it is evaluated.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59, 439-483. Also available as *CSE Technical Report 414*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/TECH414.PDF>

Foundational, not overly technical, discussion of the role that probability-based reasoning plays in assessment and assessment design.

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* (437-446). San Francisco: Morgan Kaufmann. Also available as *CSE Technical Report 518*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/TECH518.pdf>

Discussion of Markov Chain Monte Carlo estimation in a binary skills multivariate latent class model for cognitive diagnosis. Illustrated with analysis of data from Kikumi Tatsuoka's studies of mixed number subtraction.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice* (pp. 97-128). Hillsdale, NJ: Erlbaum. Also available as *CSE Technical Report 500*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/TECH500.pdf>

Discussion of assessment design issues in cases when targets of inferences about what examinees know and can do are conceived more broadly than can be observed in the context of any particular set of tasks.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Haertel, G., & Penuel, W. (2003). Improving educational assessment. In G. D. Haertel & B. Means (Eds.), *Evaluating educational technology* (pp. 149-180). New York: Teachers College Press. Also available as *CSE Technical Report 534*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved May 1, 2005 from <http://www.cse.ucla.edu/CRESST/Reports/newTR534.pdf>

Looking from the perspective of ECD at ways that assessment can be improved by developments in statistics, technology, and cognitive psychology.

Mislevy, R. J., Wilson, M. R., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. In T. Kellaghan & D. Stufflebeam (Eds.), *International Handbook of Educational Evaluation* (pp. 489-531). Dordrecht, the Netherlands: Kluwer Academic Press. Also available as *CSE Technical Report 583*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved June 6, 2005 from <http://www.cse.ucla.edu/reports/TR583.pdf>

Exploration of validity, reliability, comparability, and fairness, as viewed from the perspective of evidentiary arguments.

Williamson, D., Mislevy, R. J., & Almond, R. G. (2000). Model criticism of Bayesian networks with latent variables. In C. Boutilier & M. Goldszmidt (Eds.), *Uncertainty in artificial intelligence 16*, pp. 634-643. San Francisco: Morgan Kaufmann.

An initial investigation into model-fit indices for the use of Bayes nets in educational assessments.





Sponsor

The National Science Foundation, Grant REC-0129331

Prime Grantee

SRI International. *Center for Technology in Learning*

Subgrantees

University of Maryland

University of California, Berkeley. *Berkeley Evaluation & Assessment Research (BEAR) Center and The Full Option Science System (FOSS)*

University of Michigan. *BioKIDS*

