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

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

PADI Technical Report 9 – Draft Version

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 Design in Inquiry (PADI) project.

Acknowledgements

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 presentation 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, or ECD, for short. ECD builds on developments in fields such as expert systems (Breese et al., 1994), software design (Gamma et al., 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. While specific tasks and collections of tasks constitute one way of going about 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 (*Standards for Educational and Psychological Testing*, 1999, p. 172). 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 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 here. 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[1] (KSAs), 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 Toulmin's (1958) work on argumentation help 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

---

[1] Industrial psychologists use the phrase "knowledge, skills, or abilities", or 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.

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 representational forms like Toulmin's and Wigmore's 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: a student model, 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 kinds 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 comes with 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, for example, matrix representations of psychometric models for statisticians, and domain maps for content areas 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 (Simon, 1969; Dym, 1994). 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 time scale (as in sequential processes) for which it is possible to construct knowledge representations to support communication across subsystems as required by the overall process. While 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 Dully, 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 an heuristic for making decisions at each step in the life of a building. By employing the layers approach, activities can take place within layers which do not impact the others, yet which 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's 7-layer Open System Interconnection (OSI) Reference Model, 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: The 7 layers of Cisco's OSI]]

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. While 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 draw on Mislevy, Steinberg, and Almond (2003) and on work in the Principled Assessment Design for Inquiry (PADI) project (Mislevy & Baxter, in press). PADI is an NSF-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 take as a running example assessments from a graduate course in the foundations of assessment design, EDMS 738. 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 that guide good 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]]

[[Figure 3: Graphic representation of ECD layers]]

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. While 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 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. Just why is it that level of performance on a given task ought to 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 design 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 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, design layer, namely Domain Modeling.

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 to 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 which 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 concern yet in this layer. Rather it is the articulation of 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 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 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: Basic Toulmin diagram]]

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: Elaborated Toulmin diagram]]

PADI has adapted structures called design patterns from architecture (Alexander, Ishikawa, & Silverstein, 1977) and software engineering (Gamma, Helm, Johnson, & Vlissides, 1994; Gardner, Rush, Crist, Konitzer, & Teegarden, 1998) to help organize information from domain analysis into the form of potential assessment arguments (Mislevy, Chudowsky, Draney, Fried, Gaffney, Haertel, 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. We thus 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.]]

[[Table 3: Model elaboration design pattern.]]

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

which 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, skill, 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 focus rather 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 implies 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 chose 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 providing 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 can't 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. Since 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 towards 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 which 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, we will have in hand the assessment argument 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 variables 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 abilities, knowledge, and skills. 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, namely 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 of ECD terminology and applied ECD principles to the analyses of their chosen assessments.

While the evaluation component tells us how to characterize the salient features of any particular performance, it remains to synthesize data like this 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 nuggets 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 more as 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 the 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 chose 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 thus 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 this 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 formally in unified modeling language (UML; Booch, Rumbaugh, & Jacobson, 1999) or extended 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 which 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 which are used as substance for the Activities and Task Model Variables attributes in this template.

[[Figure 7: Graphic representation of template objects]]

[[Figure 8: Graphic representation of EDMS 738 Template]]

[[Figures 9 through 11: PADI Design System screen shots of template, activity, and task model variable objects from the EDMS 738 example]]

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. While 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 they are blueprints for specific tasks. These more specific

structures are referred to as task specifications, or task specs. While 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 pre-determined 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), has developed Question and Test Interoperability (QTI) standards to help developers share materials and

processes across assessment systems and platforms. The interested reader is referred to the IMS web site[2] for details about QTI standards. 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: Four-Process delivery Architecture]]

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

---

[2] http://www.imsglobal.org/question/index.cfm

to the evaluation procedures specified in the evidence model. Possibilities include the quality of writing, or 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.

Following 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 difference 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: Four-Process delivery Architecture--elaborated]]

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 evidence-centered design (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. A number of publications provide the interested reader with further ECD-related theoretical considerations as well as practical examples of ECD applications; these resources are identified in the annotated bibliography presented in the Appendix.

Today's 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—are wont to dominate considerations of constraints in the assessment design process. By engaging in the creation of design patterns, 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 ETS 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

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). http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

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

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. (2004). *Participatory, example-based data modeling in PADI* (PADI Technical Report 4). Menlo Park, CA: SRI International.

Joint Committee on Standards for AERA, APA, and NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

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

Kane, M. (in press). Content-related validity evidence in test development. In S. Downing & T. Haladyna (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. & Baxter, G. (in press). The case for an integrated design framework for assessing science inquiry.

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.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions*. Cambridge, MA: MIT Press.

Webb, N. (in press). Identifying content for assessing student achievement. In S. Downing & T. Haladyna (Eds.), *Handbook on test development*. Mahwah, NJ: Erlbaum.

World-Wide Web Consortium (1998). *Extensible Markup Language (XML)*,http://www.w3c.org/TR/1998/REC-xml-19980210.

Wigmore, J.H. (1937). *The science of judicial proof* (3rd Ed.). Boston: Little, Brown, & Co.

Table 1: Summary of ECD Layers

Layer	Role	Key Entities	Examples of Knowledge Representations
Domain Analysis	Gather substantive information about the domain on interest that will have direct implications for assessment, including how that information is learned and communicated.	Concepts, terminology, tools, and representations forms. Analyses of information use.	All the many and varied representational forms and symbol systems in a domain (e.g., algebraic notation, maps, content standards lists, syllabi).
Domain Modeling	Expresses 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	Expresses 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 templates.	Algebraic and graphical representations of measurement models; PADI task template object model.
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; ASCI files of item parameters.
Delivery	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 what forms as used in interactions; numerical and graphical summaries for individual and group-level reports; IMS/QTI compatible files for results.

Table 2: Design Pattern Attributes and Corresponding Assessment Argument
Components

ATTRIBUTE	DESCRIPTION	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 KSAs (knowledge/skills/attributes).	Evidence Model
Potential work products	Modes, like a written product or a spoken answer, in which students might produce evidence about KSAs (knowledge/skills/attributes).	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

Design Pattern Attribute	Comments
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 DP 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 working with an existing model.	*Students' work is bound by the concept of an existing model (or models) so their work includes an understanding 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 ♣ 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 DP focuses on establishing correspondences among models and between models and real-world situations.*
Additional Knowledge, Skills and Abilities ♣ 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.*
Potential observations ♣ 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 work products ♣ Correspondence mapping between elements or relationships of model and real-world situation ♣ Correspondence mapping between elements or relationships of overlapping models ♣ Elaborated model ♣ Written/Oral Explanation of reasoning behind elaboration	*These are examples of things that students might be asked to say, do, or construct, where their actions can provide clues about their proficiencies with model elaboration.*
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 DP 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 ♣ 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?	

Figure 1: Layers of change in buildings, adapted from *How Buildings Learn* (Brand, 1994).

	Stuff: Familiar and surface-appointments of an office or living space, such as furniture, artwork, telephones, and appliances.
	Space Plan: Layout of living or working space, including partitions, or desks and bookshelves used to delineate different spaces.
	Services: Infrastructure elements or systems such as air-conditioning, intercoms, light, power, networks. Often require substantial investment to upgrade (i.e., adding air-conditioning to existing house, or installing an additional telephone line).
	Skin: The façade of the building, both aesthetic and functional relevance (e.g., climate will play a role). Changes here are rare and expensive.

		Structure: Concrete pillars, steel girders, and wooden frames are all examples, and are fundamental. Changes here are often prohibitively expensive and substantially disruptive to normal activities housed in the building.
		Site: Location, and what the structure is built on. Could refer to land, as in sheet rock or clay, or to cultural heritage or way of thinking.

Figure 2. The 7 Layers of OSI (reproduced from [The Abdus Salam International Centre for Theoretical Physics[3]](#)).
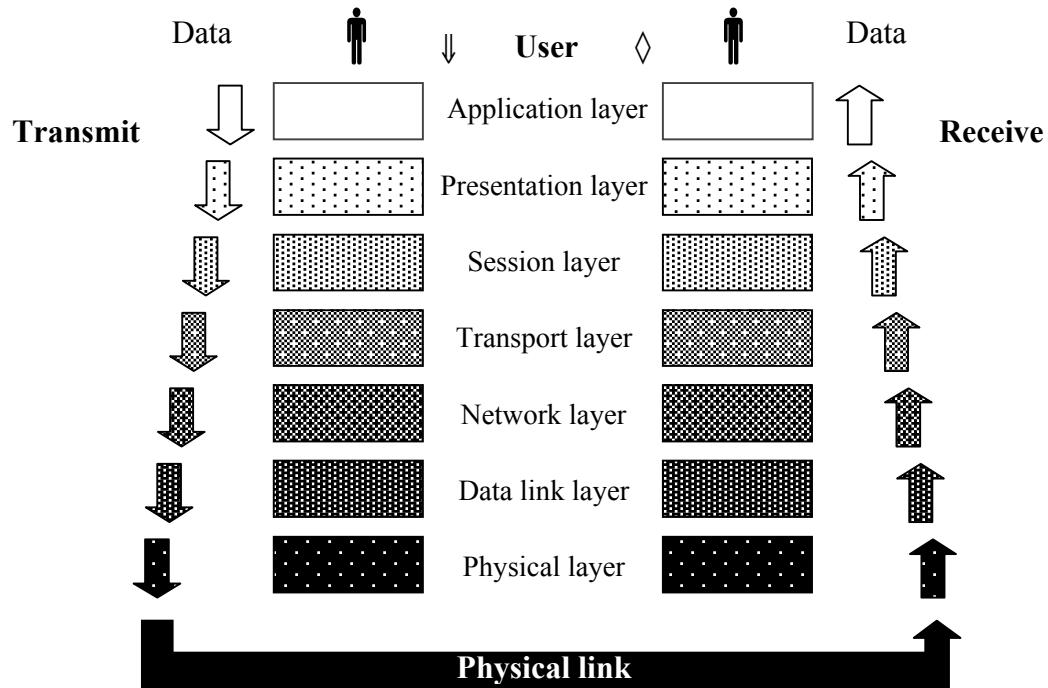
[3] http://www.ictp.trieste.it/%7Eradionet/1998_school/networking_presentation/page5.html

Domain Analysis
Domain Modeling   *[Design Patterns]*
Conceptual Assessment Framework   *[Templates and Task Specs]*
Assessment Implementation
Assessment Delivery   *[Four-Process Delivery System]*
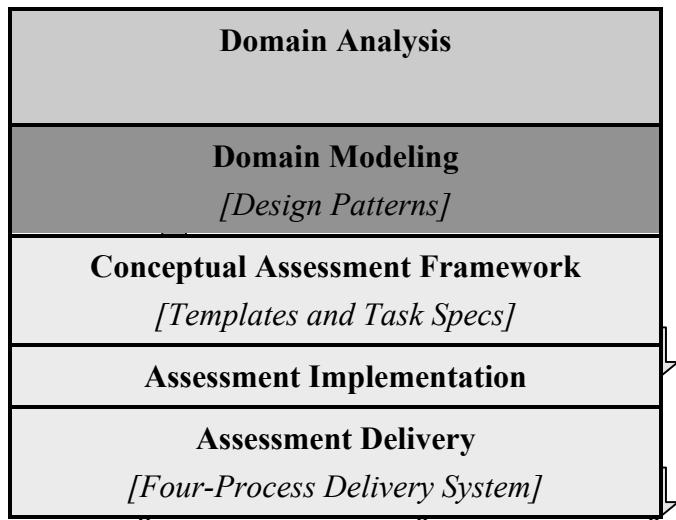
Figure 4: Toulmin's (1958) structure for arguments.
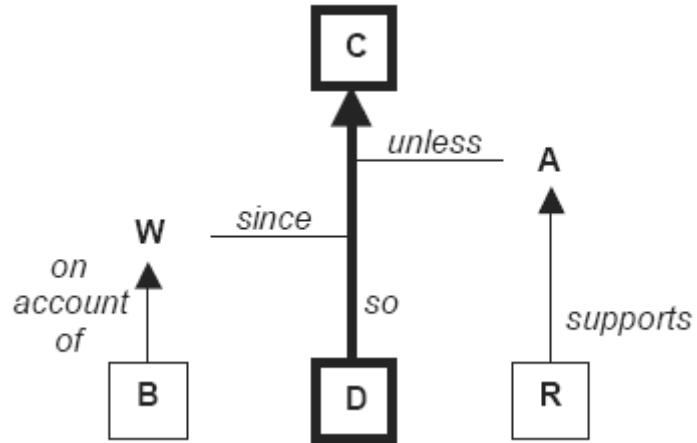
Reasoning flows from *data* (D) to *claim* (C) by justification of a *warrant* (W), which in turn is supported by *backing* (B). The inference may need to be qualified by *alternative explanations* (A), which may have *rebuttal evidence* (R) to support them.
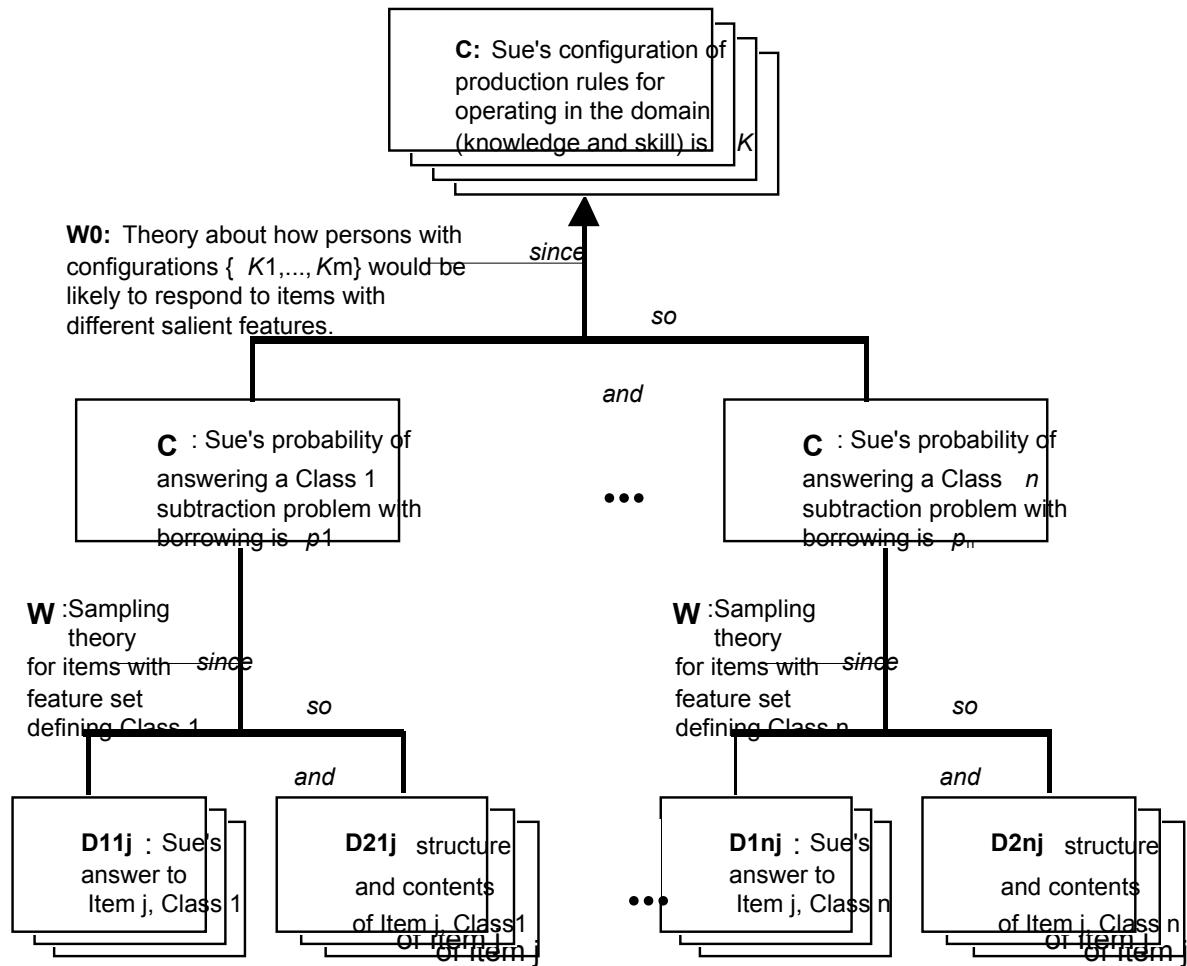
Figure 6: Conceptual Assessment Framework (CAF)

Delivery Model
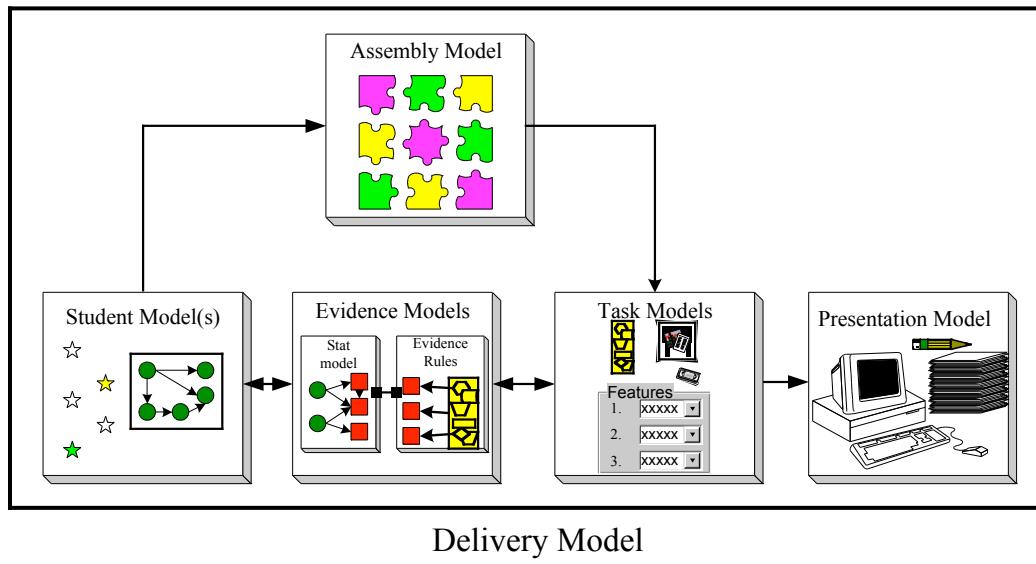
Figure 7: PADI Template Objects
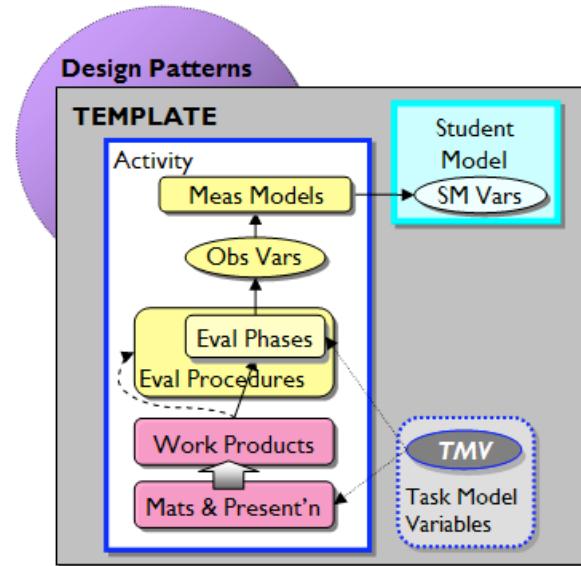
In this figure, color is used to map the objects in the PADI Design System to the ECD Models, with blue denoting the ECD Student Model, yellow the ECD Evidence Model, and pink the ECD Task Model.

Figure 8: EDMS Template Collapsed View
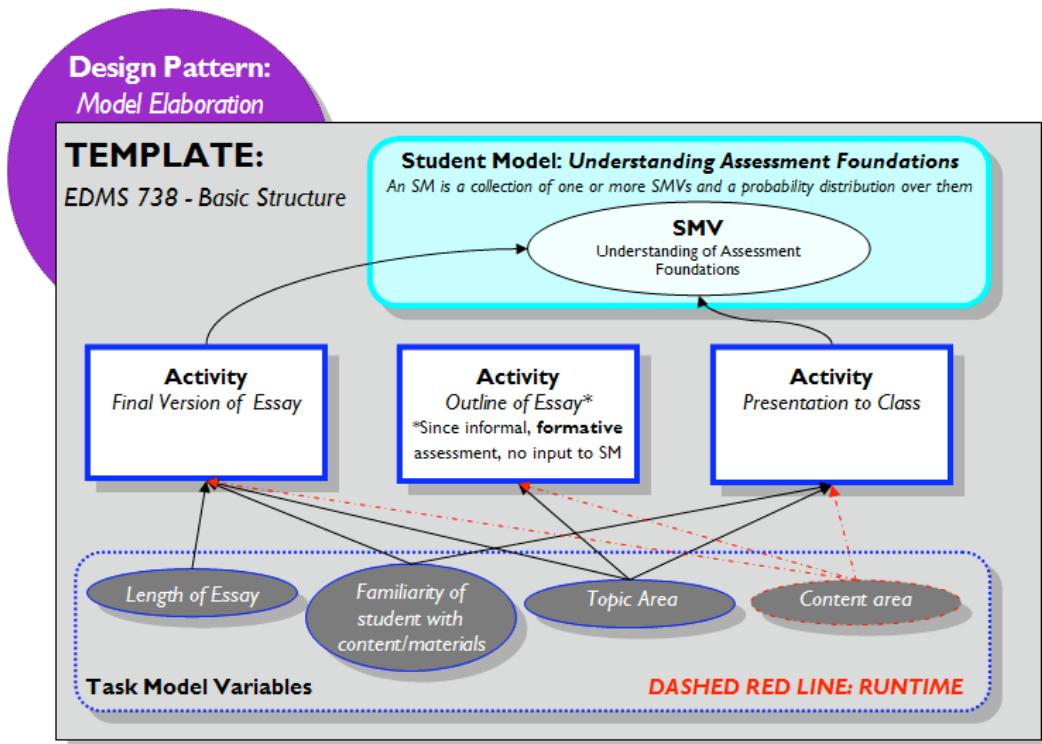
Figure 9: EDMS 738 Template in the PADI Design System

PADI

Design Patterns | Templates | Task Specifications

Education Standards | Exemplars

Student Models | Activities

Student Model Variables

Meas. Models — Observable Variables

Eval. Procedures — Evaluation Phases

Work Products

Materials & Presentation

Task Model Variables

EDMS 738 Assignments | Template 78

[Convert to Task Spec | Duplicate | Export | Delete]

Component		Value	Comment
Title		[Edit] EDMS 738 Assignments	
Summary		[Edit] 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	⊕	[Edit] Abstract Template	
Student Model Summary	⊕	[Edit] one overall summary variable of proficiency	
Student Models	⊕	[Edit] EDMS Overall Proficiency Model. Defines a univariate student model, with a continuous variable that signifies proficiency in applyin...	
Measurement Model Summary	⊕	[Edit] univariate	
Evaluation Procedures Summary	⊕	[Edit] generic rubrics	There are rubrics associated with the activity phases that can be used across specific topic areas.
Work Product Summary	⊕	[Edit] 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	⊕	[Edit]	
Task Model Variable Settings	⊕	[Edit] [View]	
Presentation Environment Requirements	⊕	[Edit] 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.	

Environment Requirements		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.
Materials and Presentation Settings	[Edit]	[View]
Activities Summary	[Edit]	1. (optional) review by instructor of outline by examinee 2. final draft 3. (optional) presentation to class
Activities	[Edit]	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.
Template-level Task Model Variables	[Edit]	Length of essay. Long or short assignments topic area. topics for essay about assessment
Tools for Examinee	[Edit]	computer with MS Word textbook, readings for course
Exemplars	[Edit]	
Educational Standards	[Edit]	
Design Patterns	[Edit]	Model elaboration. One type of problem solving that involves learning is the explanation and use of a given model.
I am a kind of	[Edit]	
These are kinds of me	[Edit]	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	[Edit]	
Online resources	[Edit]	http://www.education.u... Used in EDMS 738 Fall 2002, "Cognitive psychology and educational assessment"
References	[Edit]	
I am a part of		

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

Design Patterns | Exemplars
Education Standards
Templates | Task Specifications
Activities
Student Models | Meas. Models | Eval. Procedures | Work Products | Materials & Presentation | Task Model Variables
Student Model Variables | Observable Variables | Evaluation Phases

Final version of essay | Activity 82

[Duplicate | Export | Delete]

| Component		Value	Comment
Title	[Edit]	Final version of essay	
Summary	[Edit]	This is the final essay that is turned in for a grade.	
Measurement Models	⊕ [Edit]	Essay Grade Evidencing Proficiency. Rasch partial credit measurement fragment, modeling EDMS 738 essay grades as dependent on EDMS overa…	
Evaluation Procedures	⊕ [Edit]	Essay grading procedure.	
Work Products	⊕ [Edit]	Final Essay. Essay for final submission for grade	
Materials and Presentation	⊕ [Edit]	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	⊕ [Edit]	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	⊕ [Edit]	Content area. Specific domain content under consideration Familiarity of student with content/materials. Length of essay. Long or short assignments topic area. topics for essay about assessment	
Design Patterns	⊕ [Edit]		
Online resources	⊕ [Edit]		
References	⊕ [Edit]		
I am a part of	⊕	EDMS 738 Assignments. (Template) EDMS 738 Task Spec I - Psych and Your Assessment. (Template) EDMS 738 Task Spec II - Final Essay. (Template)	

Figure 12: 4-process architecture (adapted from …)
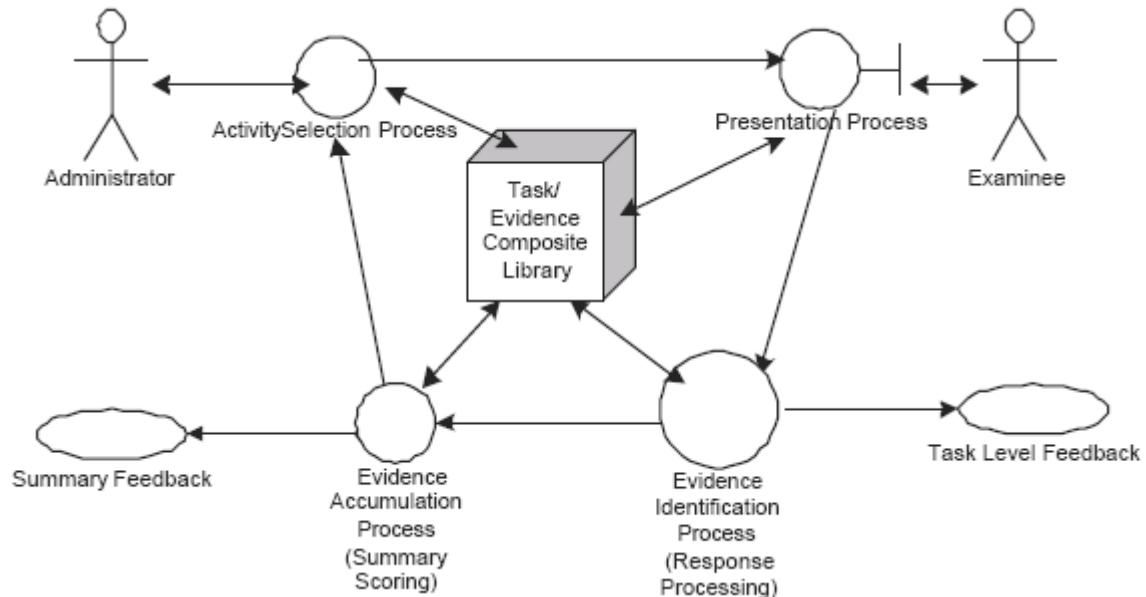
Figure 13: Processes and messages in the delivery cycle (adapted from…).