3

EVIDENCE-CENTERED DESIGN

Michelle M. Riconscente, Robert J. Mislevy and Seth Corrigan

Introduction

In this chapter we describe the principles and processes of evidence-centered design (ECD), a comprehensive assessment design framework introduced by Mislevy, Steinberg and Almond (2003). Taking its cue from developments in fields such as expert systems (Breese, Goldman & Wellman, 1994), software design (Gamma, Helm, Johnson & Vlissides, 1994) and legal argumentation (Tillers & Schum, 1991), ECD offers a method for articulating and constructing assessment arguments and the processes that instantiate them. Through a system of process layers, structures and representations, ECD facilitates communication, coherence and efficiency in assessment design and task creation.

This first section presents a guiding definition of assessment, and then introduces three concepts central to ECD: evidentiary reasoning, knowledge representations and process layers. In the remainder of the chapter, we elaborate on each layer of the process, illustrated with examples from GlassLab’s and GameDesk’s applications of ECD to educational game development. More detailed discussion of the application of ECD to simulation- and game-based assessment appears in Behrens, Mislevy, DiCerbo and Levy (2012), Mislevy et al. (2014), Shute (2011) and Shute and Wang (in press).

Defining Assessment

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.

(Messick, 1994, p. 17)

Evidence-centered design takes its cue from this construct-centered approach offered by Messick. In ECD, assessment is defined fundamentally as a chain of reasoning that links evidence to claims about learning. More specifically, assessment is the process of reasoning from the particular things people make, say or do to draw inferences about their knowledge, skills and abilities. All of the principles, structures and tools of ECD follow from this core conceptualization.
Tests have long taken familiar formats—sets of primarily constructed-response or selected-response items. However, if we conceive of assessment as the broad set of processes and instruments by which we arrive at inferences about learner proficiency, we can think of tests as any vehicle that effectively elicits and interprets evidence to draw valid inferences about a learner’s qualities (AERA/APA/NCME, 2014, p. 183). As we illustrate in this chapter, new technologies open opportunities for new forms of tests, such as simulations and digital games. ECD principles help in designing not only familiar kinds of large-scale tests but also informal classroom quizzes, formative tests, tutoring systems and performance assessments. For innovations in complex assessments, which elicit and generate evidence for “hard-to-measure” (Stecher & Hamilton, 2014) abilities and knowledge, using ECD becomes critical for organizing principled thinking about the ways that evidence is evoked and used, and for supporting communication among assessment experts and their simulation-design and game-design colleagues throughout the design and implementation process. The effort associated with implementing ECD stands to pay off further by improving the generalizability and reusability of the assessment designs we create—potentially lowering the costs of developing more novel forms of assessment (DeBarger & Riconscente, 2005).

The need for a framework like ECD has become pressing in an educational context that increasingly demands assessments capable of integrating large amounts of data and targeting complex abilities and knowledge, such as 21st-century skills (Darling-Hammond & Adamson, 2014). While advances in cognitive, psychometric and technological tools and concepts offer unprecedented possibilities for innovating the way assessments are designed and implemented, effectively leveraging these opportunities is far from straightforward. As Mislevy et al. (2003) describe, advances in assessment-related fields have had limited impact on everyday assessment practices, because the field still lacks tools for making sense of rich data for ambitious inferences (DiCerbo et al., in press). We need methods for integrating these new insights and capabilities into innovative assessment thinking and practices. This is especially true when considering that assessment design entails much more than the nuts and bolts of task authoring. A hallmark of ECD is thus to commence the assessment design process by articulating a chain of reasoning that links evidence to claims about target constructs. Only in subsequent phases are the particulars of the assessment brought to life in the machinery of tasks, rubrics, scores and the like.

**Evidentiary Reasoning and Assessment Arguments**

While the particular forms and procedures of different assessments may vary, the fundamental reasoning process that links evidence to claims does not change. To illuminate and structure this process, ECD draws heavily on ideas and terminology from Wigmore’s (1937) and Toulmin’s (1958) work on argumentation and evidentiary reasoning in the legal sector. In court cases, lawyers argue from a rich and diverse pool of observable evidence to justify a specific conclusion, or claim. Wigmore and Toulmin created graphic representations to both illustrate and facilitate evidentiary reasoning. Six fundamental elements make up an argument based on evidentiary reasoning: (1) data, (2) claim, (3) warrant, (4) backing, (5) alternative explanation and (6) rebuttal. The warrant and its backing provide the rationale or generalization for grounding the inference in the available data. They establish the credibility, relevance and strength of the evidence in relation to the target conclusions (Schum, 1994, xiii). The alternative explanations, supported by a rebuttal, describe situations or conditions that potentially weaken or even dissolve the link between the data and the proposed inference. Figure 3.1 presents Toulmin’s structure for arguments.

From an ECD perspective, educational assessment embodies these same fundamental processes. When we design and carry out assessments, our goal is to make one or more claims about student knowledge, skills or abilities (KSAs). We want these claims to be valid (Kane, this volume). This is true whether the assessment is delivered in the classroom to inform tomorrow’s lesson or in the
context of a national large-scale effort to ascertain strengths and areas in need of improvement in the formal educational system. 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 any claim, we need to obtain relevant data where criteria for their strength and relevance are determined by the warrant—what we know and value about proficiency, and what people might say or do in particular situations that could provide clues about their proficiency. Importantly, data become evidence only when their relevance to some inference is established. The same data might be good evidence for one inference but poor evidence for another (Schum, 1994). A pervasive challenge in game- and simulation-based assessment lies in transforming the vast amounts of available data into meaningful evidence for outcomes of interest. We also need to carefully consider alternative explanations for the data, as for instance when a student provides the incorrect response to a fractions problem not because of a lack of mathematics knowledge but due to limited language comprehension. The more complex and interrelated the collection of evidence and warrants, the more helpful and necessary it is to have a framework for organizing the individual and collective contributions of these elements to our claim.

As an illustration, Figure 3.2 adapts Toulmin’s and Wigmore’s representations to an assessment argument. Multiple data sources, through multiple accompanying warrants, are brought to bear on a claim about a student’s ability to organize elements of an argument from an information processing perspective: Sue completed a number of game-based tasks that require her to organize elements of an argument—supporting her ideas with evidence and backing, considering alternative positions and so forth. An information processing perspective characterizes students 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. The backing comes from both classroom experience and cognitive research (e.g., VanLehn, 1990). Patterns of responses across 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 inferences about which of the operations Sue has mastered and which she has not.
Assessment design draws on specialized knowledge across diverse areas of expertise. This is all the more true as supporting technologies become more advanced, and applications such as digital simulations and games are leveraged for their assessment potential. For example, the team at GlassLab Games that develops game-based assessments includes assessment designers, learning designers, teachers, database engineers, software designers, cognitive psychologists, game designers and psychometricians (Mislevy et al., 2014). Supports for effective communication within these multidisciplinary teams are crucial. Across each field, there exist distinct language and methods that have evolved to solve specific problems. In the process of designing game-based and other assessments, specialists within each relevant field will engage in conversations and processes requiring varying forms of specialized knowledge. The outcomes of those conversations will eventually be communicated to a broader group and integrated with considerations from other topics into the overarching assessment argument. Consequently, there is a need for a common language and a common framework to orchestrate and integrate the contributions of diverse areas of expertise into a coherent assessment.

Related to the need for a common language are shared knowledge representations (Markman, 1998). Information, in order to be useful, must always be represented in some form. Good representations—such as Toulmin diagrams—capture the important features of information in a form that facilitates reasoning with and applications of that information. The better these representations are aligned to the purpose the information is to serve, the more powerful and effective they will be. Knowledge representations are important in educational assessment since different representations of the information will be optimal for different people and different processes (Mislevy et al., 2010). We have found that a variety of knowledge representations—including design patterns, psychometric models...
and task templates—are essential for applying ECD to construct a solid underlying assessment argument. These representations have the additional benefit of helping heterogenous teams to understand the structure and process of assessment. In subsequent sections of this chapter, we introduce several knowledge representations that have evolved in the context of applications of ECD to solve a variety of assessment design challenges.

A Layered Approach

In addition to evidentiary reasoning and knowledge representations, ECD leverages the concept of layers to support intrafield investigations while simultaneously providing structures that facilitate communication across various kinds of expertise, as each contributes to the assessment argument (Dym, 1994; Simon, 1969). Layering is an effective strategy for tackling large, complex processes composed of many distinct yet related tasks and topics. In this approach, the overall process is segmented into coherent layers, each with its own characteristic tasks and processes. Work is carried out within each layer independently of the others, and at key points of the process, outcomes are passed from one layer to another through knowledge representations. The layers are related to one another by characteristics such as time scale or sequences (as in sequential processes), for which it is possible to construct knowledge representations to support communication across layers 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.

As we describe in detail in the second section, each layer of ECD is defined by a set of goals, tasks, associated expertise and knowledge representations. Within each layer there are interrelated conceptual, structural and operational considerations to coordinate. Understanding relationships within layers clarifies decision points and issues involved in making them. The depictions of layers and various representations within layers discussed in this chapter draw on Mislevy et al. (2003) and on the Principled Assessment Design for Inquiry (PADI) project (Baxter & Mislevy, 2004; Mislevy & Haertel, 2006), as well as on GlassLab’s process for creating game-based assessments (Mislevy et al., 2014) and GameDesk’s game development process (Riconscente & Vattel, 2013).

Examples and Applications

Since its inception, ECD has been applied to an increasingly diverse range of projects, from performance-based tasks to assess science inquiry, to adaptive assessments that accommodate learners with specific physical interaction constraints, to game-based assessments and simulations for learning. ECD has been used both to forward-engineer assessments and to work from existing effective assessments to “reverse-engineer” them into the knowledge representations of ECD so that they can be refined and reused to generate more assessments. To showcase ways that careful application of ECD aids in the design of digital simulation- and game-based assessments, in this edition of the Handbook we present examples drawn from game-based assessments developed by GlassLab and GameDesk to treat complex abilities and knowledge.

The ECD Layers

This section walks through the ECD layers, noting the kinds of work that take place within and across layers, and offers examples of knowledge representations in each layer, as summarized in Table 3.1. Since veterans of test development will find more familiar terms and concepts in the layers closest to task creation and implementation, our discussion focuses on the initial layers in which the assessment argument is articulated and elaborated. Although it usually remains in the background, a coherent, targeted assessment argument implicitly guides the design choices that good task developers make.
Domain Analysis

The goal of domain analysis is to gather substantive information about the target domain, and to identify the KSAs about which we will make assessment claims (AERA/APA/NCME, 2014, p. 76). This information includes the content, concepts, terminology, tools and representational forms that people utilize within the domain. It may include the situations in which people use declarative, procedural, strategic and social knowledge as they interact with others and the environment. It may entail task surveys of how often people encounter various situations and what kinds of knowledge demands are important or frequent, and cognitive analyses of how people use their knowledge. Through analysis of real-world tasks, practice analysis extracts features that are important for carrying out the responsibilities of a certain job (Raymond & Neustel, 2006), which in turn inform the kinds of student KSAs to assess.

Domain analysis also includes, at least implicitly, one or more conceptions of the nature of knowledge in the targeted domain (Perie & Huff, this volume). For example, mathematics can be viewed...
through the lenses of the behavioral, information processing or sociocultural perspectives (Greeno, Collins & Resnick, 1997). A strict behaviorist perspective would concentrate on procedures for solving problems in various classes—possibly complex procedures, but conceived of assemblages of stimulus-response bonds. An information processing perspective would emphasize the cognitive processes underlying acquisition of mathematics knowledge, and seek to identify reasoning patterns that indicate students are on track. 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, an assessment designer would design situations to observe students act in ways that gave evidence for the kinds of inferences that were being targeted. Rather different tasks, evaluation procedures and reports would emerge. Since the psychological perspective fundamentally drives the choice of content taught and assessed, expectations for student KSAs and ways of assessing progress, it should be clearly articulated and referenced throughout the assessment design process. A mismatch in psychological perspectives at different stages results in substantially less informative assessment.

Assessment design can start from a variety of points, such as claims about student proficiency (e.g., “verbal ability”), 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 language assessment tasks), or the qualities of work at increasing levels of proficiency (e.g., learning progressions, as in West et al., 2012). Although the target inferences associated with different starting points will vary, all eventually require a coherent chain of observations in order to arrive at valid claims (Kane, this volume; AERA/APA/NCME, 2014, p. 21). It is worth noting that for this reason, a variety of methods for identifying evidence, including educational data mining, can be perfectly compatible with an ECD approach to assessment design.

Organizing categories help designers shape information about a domain and an assessment purpose in ways that subsequently can be easily translated into assessment arguments. These categories include valued work, task features, representational forms, performance outcomes, valued knowledge, knowledge structure and relationships, and knowledge-task relationships. Each category looks back toward the domain to capture features that make sense to teachers, domain experts and researchers in the domain. At the same time, they look forward, organizing information in ways that facilitate domain modeling, the next layer.

We identify valued work by examining real-world situations in which people engage in the behaviors and utilize the knowledge key to the domain. From these situations we can ascertain the kinds of tasks appropriate for assessment, as well as features of performances that are important to capture in assessment. Salient features of the situations in which this valued work can be observed are task features. The assessment designer will manipulate task features to focus evidence, stress different aspects of knowledge and constrain alternative explanations for performance.

In any domain, people use a variety of representational forms. Learning how to use representational forms to characterize situations, solve problems, transform data and communicate with others is central to developing proficiency. Musical notation, for example, has been developed for representing compositions, with some universals and some instrument-specific features. Not only is much of the knowledge in domains built into these representations, but also they are used to present information and capture thinking in assessment tasks (Gitomer & Steinberg, 1999).

Performance outcomes indicate the ways we recognize, from what they have said or done, students’ understandings. These characteristics form the criteria that will be used to craft rubrics or scoring algorithms. Characteristics of the knowledge, or content, of a domain will also be central to assessment design. These are referred to as valued knowledge. Curriculum materials, textbooks and concept maps are examples of sources of valued knowledge, as are state and professional standards documents.

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 here. Finally, we need to explicate knowledge-task relationships, or how features
of situations and tasks interact with knowledge. These help us identify task features that reveal differences in examinees’ understandings.

The domain analysis layer is furthest from the concrete tasks we ultimately generate in assessment design. But the thinking along the lines sketched earlier underscores the importance of this layer in the overall process, to build validity into assessment outcomes from the start (AERA/APA/NCME, 2014, p. 11).

A careful domain analysis was central to the design process GameDesk used to create AERO, a 3-D interactive game in which students “become” an albatross to learn and demonstrate core concepts regarding the physics of flight (Photo 3.1). Identification of key knowledge, as well as common misconceptions, helped focus the game interactions and assessment data capture. For instance, in order to understand lift in the context of flight, students must understand that gravity always points not just “down” but specifically toward the center of the earth. Similarly, understanding the relationship between lift and the wings’ rotation is necessary for maintaining flight. Domain analysis also revealed shortcomings in traditional approaches to teaching this topic, as well as the importance of dynamic rather than static force diagrams that students could interact with and observe in real time during flight. Visualizing the continual cause-and-effect relationship between their actions and the resulting effect on the force vectors became a key game feature to support learning, with those same representations integrated into assessment sections of the game experience.

**Domain Modeling**

In the domain modeling layer, we harvest and organize the results of the domain analysis process to articulate, in narrative form, an assessment argument that connects observations of students’ actions in various situations to inferences about what they know or can do. Whereas contributions from content...
and instructional experts are the foundation of domain analysis, the assessment designer plays a more prominent role in domain modeling.

Toulmin’s general structure for arguments presented in the previous section provides a helpful starting point for identifying in broad strokes the claims, data and warrants that will make up the line of reasoning for the assessment argument being created (Figure 3.1). In all assessments, but especially in those with complex, interactive performances, we expect pilot testing and think-aloud trials with early prototypes to provide valuable insights that will circle back to add forms of evidence or sharpen arguments. Data mining of log files from game and simulation data, for example, can lead to additional features of student performances to be recognized and captured, as well as suggest improvements in task features and directives to better elicit evidence.

Domain modeling can be carried out using any of a variety of knowledge representations. We have found that simple structures called design patterns, which originated in architecture (Alexander, Ishikawa & Silverstein, 1977) and software engineering (Gamma et al., 1994), are helpful for organizing information from domain analysis into the form of potential assessment arguments (PADI, 2003). Because the structure of the design pattern follows 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. The assessment structure is thus provided by the design pattern, while the assessment substance is determined by the assessment designer (Mislevy, 2003).

Table 3.2 shows the attributes of a sample design pattern and their connection to the assessment argument for the Mars Generation One game-based assessment developed by GlassLab. Set in the year 2054, Mars Generation One (MGO) unfolds at the first human settlement on Mars, where citizens settle their differences and make important policy decisions by sending robot assistants—“argubots”—into argument duels. Players learn to equip their argubots with valid arguments through a series of missions that require them to gather evidence, build digital claim-cores that pair

![Photo 3.2 Mars Generation One screenshot.](image-url)
relevant and supporting evidence to claims, and participate in debates by evaluating and critiquing
the arguments of others, all while defending their own arguments. As students play, the game gathers
data on their in-game performances, to support claims about students’ ability to develop, evaluate
and critique arguments.

Design patterns are intentionally broad, narrative and nontechnical. Centered around a particular
KSA, a design pattern allows for a variety of approaches that can be used to gather evidence about
that knowledge or skill, organized in such a way as to lead toward the more technical work of design-
ing particular tasks. There exist many examples. Among the several design patterns the PADI project,
out of SRI International, has developed for use in assessing science inquiry, for example, are those
for model formation and model revision (Mislevy, Riconscente & Rutstein, 2009) and experimental
investigation (Colker et al., 2010). Utilizing ECD to create design patterns such as these stands to
make the resulting assessment-based claims more generalizable. Use of ECD also stands to make the
design process more generative in the sense that many complex assessments can be generated from a
single design pattern.

Title and summary slots summarize a design pattern’s purpose and basic idea. The rationale slot
articulates the warrant that connects the target inferences and the kinds of tasks and evidence that
support them. Focal KSAs come from the valued knowledge identified in domain analysis, and
indicate the primary target of the design pattern (and the assessments to be generated). Focal as well
as additional KSAs that might also be required are cast in terms of the student, since inferences will

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Assessment argument component</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rationale</strong></td>
<td>Provide a rationale for linkages between the potential observations and focal KSAs</td>
<td>Warrant (underlying)</td>
</tr>
<tr>
<td><strong>Focal knowledge, skills and abilities</strong></td>
<td>The primary knowledge/skill/abilities targeted by this design pattern</td>
<td>Student model</td>
</tr>
<tr>
<td><strong>Additional knowledge, skills and abilities</strong></td>
<td>Other knowledge/skills/abilities that may be required to complete tasks generated with this design pattern, but which are not the target of the assessment</td>
<td>Student model</td>
</tr>
<tr>
<td><strong>Potential observations</strong></td>
<td>Observable qualities of student work products that would give evidence about the student’s proficiency with respect to the KSAs (knowledge/skills/attributes)</td>
<td>Evidence model</td>
</tr>
<tr>
<td><strong>Potential work products</strong></td>
<td>Modes, like a written product or a spoken answer, in which students might produce evidence about KSAs (knowledge/skills/attributes)</td>
<td>Task model</td>
</tr>
<tr>
<td><strong>Characteristic features</strong></td>
<td>Aspects of assessment situations that are likely to evoke the desired evidence and that are assumed to be conditions for all assessments that will be generated by this design pattern</td>
<td>Task model</td>
</tr>
<tr>
<td><strong>Variable features</strong></td>
<td>Aspects of assessment situations that can be varied in order to shift difficulty or focus, including aspects that may be defined by the student, such as the specific topic of an essay or the software tool used to create a presentation</td>
<td>Task model</td>
</tr>
</tbody>
</table>
Table 3.3 Argumentation Design Pattern for Mars Generation One

<table>
<thead>
<tr>
<th>Design pattern attribute</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Summary</strong></td>
<td>A central element of successfully participating in arguments is the ability to support one’s claims with evidence. This design pattern emphasizes two aspects of this ability: (1) identifying evidence that is relevant and supports a given claim; (2) supporting claims with evidence that is consistent with the intended argumentation scheme.</td>
</tr>
<tr>
<td><strong>Rationale</strong></td>
<td>There are many possible argumentation schemes. Argument by authority, observation, example and consequence will be treated in Mars Generation One.</td>
</tr>
<tr>
<td><strong>Focal knowledge, skills and abilities</strong></td>
<td>These are the KSAs that we intend to make inferences about. Designers should represent here important elements of the domain so that assessing them is worthwhile.</td>
</tr>
<tr>
<td>Organizing elements of an argument</td>
<td>According to the designer’s purposes, tasks may stress or minimize demand for KSAs other than the focal KSAs. This can include content knowledge, familiarity with the task type and other aspects of the activity engaged in during the assessment.</td>
</tr>
<tr>
<td>Organizing elements of an argument</td>
<td>These are aspects of things that students might say, do or construct in situations that call for argumentation. They are meant to stimulate thinking about the observable variables the designer might choose to define for assessment tasks addressing model elaboration.</td>
</tr>
<tr>
<td><strong>Additional knowledge, skills and abilities</strong></td>
<td>These are examples of things that students might be asked to say, do or construct that would provide clues about their proficiencies with argumentation.</td>
</tr>
<tr>
<td>- Identify evidence that is relevant and that supports the given claim.</td>
<td>Any task concerning organization of elements of an argument generated in accordance with this Design Pattern (DP) will indicate the set of claims, evidence and schemes the player will have to work with, along with the debate topic(s) and any additional information players will access through the nonplayer characters in the game.</td>
</tr>
<tr>
<td>- Choose evidence that is consistent with an intended argumentation scheme.</td>
<td></td>
</tr>
<tr>
<td>- Evaluate and correct claim-evidence pairs that do not exhibit coherence between the claim, evidence and/or the intended scheme.</td>
<td></td>
</tr>
<tr>
<td><strong>Potential observations</strong></td>
<td></td>
</tr>
<tr>
<td>- Relevance of the evidence for the given claim;</td>
<td></td>
</tr>
<tr>
<td>- Support status of the evidence for the given claim;</td>
<td></td>
</tr>
<tr>
<td>- Consistency between the intended argument scheme and the type of evidence selected;</td>
<td></td>
</tr>
<tr>
<td>- Self-correction when claim-evidence pairs are assembled that lack coherence with regard to relevance, support status and/or scheme.</td>
<td></td>
</tr>
<tr>
<td><strong>Potential work products</strong></td>
<td></td>
</tr>
<tr>
<td>- Review of claims and evidence;</td>
<td></td>
</tr>
<tr>
<td>- Selection of claim-evidence pairs;</td>
<td></td>
</tr>
<tr>
<td>- Revision of claim-evidence pairs before feedback;</td>
<td></td>
</tr>
<tr>
<td>- Revision of claim-evidence pairs after feedback;</td>
<td></td>
</tr>
<tr>
<td>- Evaluations of others’ claim-evidence pairs.</td>
<td></td>
</tr>
<tr>
<td><strong>Characteristic features</strong></td>
<td></td>
</tr>
<tr>
<td>- Argument topic with background information provided through the narrative and characters within the game;</td>
<td></td>
</tr>
<tr>
<td>- Multiple argumentation schemes, claims and pieces of evidence to choose from;</td>
<td></td>
</tr>
<tr>
<td>- Feedback and potential for revision of assembled claim-evidence pairs.</td>
<td></td>
</tr>
</tbody>
</table>
Evidence-Centered Design

<table>
<thead>
<tr>
<th>Design pattern attribute</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable features</strong></td>
<td>These are features of the tasks that can be manipulated to better control their difficulty level.</td>
</tr>
<tr>
<td>- Use of rhetorical markers making more or less apparent the link between the evidence type and the scheme type;</td>
<td></td>
</tr>
<tr>
<td>- Reading level of the text associated with the claim and evidence;</td>
<td></td>
</tr>
<tr>
<td>- Number of claim, evidence and scheme choices.</td>
<td></td>
</tr>
</tbody>
</table>

Concern the extent to which the student evidences them. The designer considers which KSAs are appropriate to assume, which to measure jointly and which to avoid, in order to serve the purpose of the assessment. This is accomplished by making design choices about the variable features of tasks, as discussed ahead.

In Mars Generation One, focal KSAs include creating and critiquing arguments in the form of argubots—robots that players build to deliver an argument in the game’s argument duels. Understanding the content area of the given debate, understanding how to build the argubots using the game’s interface and understanding how to operate the argubots in an argument duel are ancillary but necessary additional KSAs. The importance of the additional KSAs becomes clear when we consider what can be inferred from a student’s efforts to complete a task. Students’ content knowledge and their skills in using the game systems themselves stand to affect the quality of their responses. Noting where these additional KSAs may be required and minimizing their impact in our designs help us rule out explanations for poor responses that are based on knowledge or skills that the task requires other than the targeted, focal KSAs—sources of what Messick (1989) called construct-irrelevant variance.

Potential work products are all the things students produce—whether things they say, do or make—that we expect to hold clues about the focal KSAs. However, it is not these artifacts themselves that are the evidence for the assessment; rather it is their qualities that actually inform claims about students’ KSAs. Therefore design patterns also include a slot for potential observations, where the assessment designer articulates the particular aspects of work products that will constitute evidence for the focal KSA. Potential observations describe the qualities of work products that matter for the desired claims (e.g., “number of . . .”, “quality of . . .”, “level of . . .”, “kind of . . .”). Work products are translated into observations using potential rubrics, which identify techniques that could be used or adapted to evaluate (i.e., “score”) work products, thereby quantifying or associating values with the observations to help answer the question “To what extent does this work product meet the intended criteria?” 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 the ability to write a coherent essay, then the potential observations will attend to aspects of the work product such as the degree to which appropriate grammar is used, not the technical quality of the explanation of the process. If the focal KSA is knowledge of chemical processes, rubrics might focus instead on the accuracy of the processes described.

In the context of digitally administered assessments, information can be gathered about the processes students enact when completing the given task or challenge. Hence, these too can be used as sources of evidence when there are good reasons for believing the target aspects of the response process indicate how much or little someone knows or can do. Where it is possible to capture and score data regarding student response processes, new sources of evidence become available to the assessment designer, beyond typical work products that reveal only outcomes of a larger process.
With characteristic features and variable features, the designer specifies aspects of the situation in which the work products are produced. Characteristic implies that generally all tasks bear these features in some form, in order to support inferences about the focal KSA. Variable features are aspects of the task environment that the designer can implement in different ways. Within the constraints of the characteristic features, different configurations of variable features allow a designer to provide evidence about the focal KSA, but they can also influence the level of difficulty and the degree of confounding with other knowledge, facilitate gathering more or less evidence at lesser or greater costs and so on.

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 patterns. For example, “Problem Solving” is linked to more specific design patterns for “Solving Well-Defined Problems” and “Solving Ill-Defined Problems.” The former can provide better evidence about carrying out problem-solving procedures, but at the cost of missing how students conceptualize problems. The latter 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.

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 and described in the next section.

**Conceptual Assessment Framework**

The structures in this third layer in the ECD approach to assessment design once again express an assessment argument, but they move away from the narrative form of domain modeling and 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 (Riconscente, Mislevy, Hamel & PADI Research Group, 2005). 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 content and settings. The discussion ahead uses examples from PADI, but similar work on task modeling frameworks has been carried out by Chung et al. (2008), Embretson (1998), Luecht (2003) and others.

The CAF is machinery for generating assessment blueprints, by means of representations that coordinate the substantive, statistical and operational aspects of an assessment. Design decisions here give concrete shape to an assessment. These decisions include the statistical models, the materials that characterize the student work environment and the procedures for evaluating students’ work. The CAF layer expresses the assessment argument in operational terms, primed to generate tasks and attendant processes that inform the target inferences about student proficiency.

The CAF, sketched in Figure 3.3, 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 psychometric model, which can be as simple as an overall score across tasks or as complex as a multivariate item response theory or cognitive diagnostic model (Mislevy et al., 2014).

The CAF task model lays 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 may be
Evidence-Centered Design

identified in design patterns, which link potential work products to the KSAs. Each may have its own strengths and weaknesses. Choices among them and specific forms are now chosen to fit the purposes, the resources and the context of the particular assessment that is being designed. These more specific forms are expressed in the CAF evidence model. Marshaling multiple tasks into an assessment is coordinated by the assembly model in fixed-form and computer-adaptive tests (van der Linden & Glas, 2010). In an interactive assessment, such as a simulation or game, the assembly model ultimately takes the form of a finite state machine that specifies which challenges or game conditions are presented on the basis of the state of the simulation or game, the player's previous actions and estimates of the player's current ability with regard to the targeted KSAs (AERA/APA/NCME, 2014, p. 188).

Student Model: What Are We Measuring?

Domain analysis and domain modeling describe target inference in the form of narratives about content and student KSAs. It is not possible to observe student proficiencies directly; they must instead be inferred from incomplete evidence, as the handful of things that students say, do or make. The CAF lays out the statistical machinery for making inferences about student proficiencies, which can be expressed in terms of probability distributions over a single variable or set of variables.

In the simplest case, where a single proficiency is of interest, the student model would contain a single student model variable and students could be characterized in terms of the proportion of a domain of tasks they are likely to respond to 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 would be used to express the level of ability that is most likely for a given student.

GlassLab's game-based assessment SimCityEDU, for example, focuses on a single proficiency—systems thinking—and provides a simple example in which students are characterized in terms of a single student model. In SimCityEDU, players take the role of mayors and are responsible for simultaneously solving their city's pollution, energy and economic problems in four game-based challenges (Photo 3.3). The game is designed to assess players' facility with systems thinking by gathering evidence regarding the extent to which students identify and act upon the multiple independent variables impacting their city.
Photo 3.3  SimCityEDU screenshot.
Mars Generation One, on the other hand, offers a good example of a more complex case. There, the student model employs three different student model variables: identifying evidence for arguments, organizing evidence and evaluating arguments. As a result, a more complex multivariate probability distribution is needed to describe students’ current ability in argumentation.

**Evidence Model: How Do We Measure It?**

There are two components to the evidence model. The first concerns the qualities of the work products students have produced—for example, completeness, 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” in the context of traditional assessment items. A student’s responses across the given assessment make up the “response vector.” In the context of simulations and games, the idea is similar, as students carry out processes and generate work products in the course of their interaction with the simulation or game that can also be recorded and scored in a set of steps not much different from those of traditional assessments.

In both cases, evaluation procedures specify how the values associated with the observable variables are to be determined from students’ work products. Examples of evaluation procedures are answer keys, scoring rubrics with examples, and automated scoring procedures in computer-based games and simulation tasks. In addition, several features of a single work product may be important for inference, in which case evaluation procedures must produce values of multiple observable variables that are all associated with the same work product(s). This is true for both the SimCityEDU and the Mars Generation One game-based assessments. In one SimCityEDU challenge, for example, players are tasked with reducing pollution while simultaneously increasing the supply of power in a city dominated by coal-burning power plants. In that case the final levels of air pollution and the amount of power produced in the player’s city become important observations that are scored and then used as evidence for claims about the player’s level of ability in systems thinking.

The second part concerns the measurement model. 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 set of responses 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), cognitive diagnosis models and latent class models (e.g., for mastery testing). More common is the informal approximation of taking weighted or unweighted scores over items, which can suffice when all items contribute relatively independent nuggets of evidence about the same targeted proficiency.

**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 log file from a simulation task? What materials will be necessary as prompts? 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? If it is, then the setting should refrain from providing this information so that the task will call upon the students’ knowledge in this regard. If not, then providing open-book problems or formula sheets is appropriate, so as to focus evidence on using formulas in practical situations. The claims about students we wish to make shape the choices of task features—both those established in advance and those determined during implementation—for instance, the particular values of numbers in dynamically generated mathematics problems.

The preceding paragraphs describe task models in traditional assessments, which are either constructed by the designer (e.g., fixed-form tests) or assembled in accordance with an item selection strategy (e.g., in computerized adaptive tests). Simulation- and game-based assessments differ in that students may be able to make different choices while solving a problem or carrying out an investigation. There may be certain activities that yield work products and are required of all students. The previous discussion of task models holds for these. But in other cases, it becomes necessary to recognize situations that students work themselves into as instances of paradigmatic situations. The task model is then a description of such situations, and includes its key features and what is then to be looked for in performance (see Mislevy & Gitomer, 1996, for an example). Additional observational variables can be defined after initial implementation of an assessment, as discovered through data mining efforts. Much is to be gained from data mining when initial work in domain analysis has led to the design of simulation environments that maximize the possibility of detecting construct-relevant patterns of actions and reducing construct-irrelevant features of both situations and student choices (e.g., Gobert, Sao Pedro, Baker, Toto & Montalvo, 2012; Shute, Ventura, Bauer & Zapata-Rivera, 2009).

In GameDesk’s Geomoto, an embodied game about plate tectonics, players are challenged to create specific geographic features—such as earthquakes, volcanoes and convergent boundaries. In order to serve as an effective assessment of student’s understanding, the task model included several characteristic features informed by the domain analysis as well as iterative testing with students. For instance, the game challenges explicitly target several misconceptions related to terminology and the concepts they represent in the domain of plate tectonics. Since textbook illustrations were found lacking in their ability to offer students a sense of scale, the game includes visualizations and interactions to support student understanding that tectonic plates are immensely large, and move at exceptionally slow speeds. The task model developed for Geomoto entailed creating a variety of game challenges, each setting up “sting operations” to check whether students’ behaviors reflected understanding of the essential concepts and processes involved in convergent, divergent and transform plate boundaries and the resulting phenomenon of rift valleys, subduction zones, volcanoes, island chains and earthquakes.

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

A single piece of evidence is rarely sufficient to sustain a claim about student KSAs. 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.

The assembly model may need to be defined at a coarser grain-size for simulation- and game-based assessments. As noted earlier, it may not be a matter of selecting tasks beforehand to administer, but recognizing situations as instances of task models. Test assembly in this context corresponds to rules in the state machine that governs how ongoing situations adapt to students’ actions (Mislevy et al., 2014). For example, a challenging complication could be introduced into a computer-based patient management task only if the student is performing well.

Sample Knowledge Representations

PADI is just one of any number of systems that could be constructed as a vehicle for implementing the work of the assembly layer. The PADI project has developed structures called templates (Riconscente et al., 2005) for this layer. 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 Extensible Markup Language (XML; World-Wide Web Consortium, 1998), or in a more interactive format as web pages in the PADI design system. Within such a 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. Figure 3.4 is a generic representation of the objects in a PADI template.
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 in the case of traditional assessments. In the case of digital simulations and games, the designer is positioned to generate multiple tasks or challenges. Because they were generated from ECD, the tasks each embody a shared rationale and assessment argument in spite of possibly presenting differences in their surface features. 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; see Gierl & Lai, this volume, for a discussion on automated item generation).

An online design system such as PADI makes it possible to automate some design processes (Mislevy et al., 2010). For example, templates can be used as schemas to generate 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. This idea in fact characterizes simulation- and game-based assessment, in that the presentation process (discussed ahead) contains a library of situation-construction elements and rules to assemble them during students’ interaction with the simulation or game.

Assessment Delivery

The preceding design layers analyze a domain to determine what KSAs are of interest, and how you know them when you see them; 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. But 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, which 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. Attention here focuses on the conceptual model of the assessment delivery layer—namely, the four-process architecture for assessment delivery shown in Figure 3.5 (Almond, Steinberg & Mislevy, 2002).

Assessment operations 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 Mars Generation One, students pass through a series of training modules in order to level up in the game and gain access to new types of argubots. In the course of that training, additional challenges are presented if the student’s success rate does not meet a specified threshold.

The activity selection process typically sends instructions about presenting the item to the presentation process, which takes care of presenting the item or challenge to the student, in accordance with materials and instructions laid out in the task model. The presentation process also collects responses for scoring and analysis—that is, the work product(s). The work product may be the letter corresponding
Figure 3.5  Processes and messages in the delivery cycle.
to a multiple-choice option, or it may be a wealth of information including traces of students’ pathways navigated through the game or simulation, final responses or choices, notes made in the course of the activity and total time spent. In SimCityEDU, for example, click stream data is captured, describing each building the student destroys, each parcel of land that is rezoned, the placement of power plants and when the student accesses one or more of the game’s maps of the city, among others.

In such simulation- or game-based assessments, the rules for selecting or adapting activities are implemented in the finite state machine that governs the system’s interactions with the student more comprehensively, and alerts the other processes when actions are needed (e.g., when to present in-game feedback to the student, or when the evidence accumulation process must update the student model, so that an interim report can be generated for the teacher).

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. Examples 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 or features of the work product 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 student 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 can then inform 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—although, again, in the case of digitally based assessments, some of this can be carried out through use of finite state machines without yet applying one or more statistical or psychometric models.

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 3.5. 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 (from the evaluation specifications in the evidence model). Specifications for integrating the evidence into an accumulating student report 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 3.5 shows 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).

**Conclusion**

This chapter viewed assessment design as the development of an assessment argument, facilitated by the evidence-centered design 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 and game-based assessments by GlassLab and GameDesk.

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 nuts and bolts addressed in the 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. Off-the-shelf (or off-the-web) supports for implementing the particulars of the processes described herein are beginning to become 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 chapter provide a useful framework for new designers and a useful supplement to experienced ones.

The value 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 reusable 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 simulation- and game-based assessment. 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.

**Notes**

1. Industrial psychologists use the phrase “knowledge, skills or abilities,” or KSAs, to refer to the targets of the inferences they draw. We apply the term 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.

**References**


Michelle M. Riconscente et al.


62


