

Implications of Evidence-Centered Design for Educational Testing

Robert J. Mislevy, *University of Maryland*
Geneva D. Haertel, *SRI International*

Evidence-centered assessment design (ECD) provides language, concepts, and knowledge representations for designing and delivering educational assessments, all organized around the evidentiary argument an assessment is meant to embody. This article describes ECD in terms of layers for analyzing domains, laying out arguments, creating schemas for operational elements such as tasks and measurement models, implementing the assessment, and carrying out the operational processes. We argue that this framework helps designers take advantage of developments from measurement, technology, cognitive psychology, and learning in the domains. Examples of ECD tools and applications are drawn from the Principled Assessment Design for Inquiry (PADI) project. Attention is given to implications for large-scale tests such as state accountability measures, with a special eye for computer-based simulation tasks.

Keywords: assessment design, delivery system, evidence-centered design, PADI

These are heady times in the world of educational assessment—days of urgent demands, unprecedented opportunities, and tantalizing challenges. Demands for consequential tests in schools and states, at larger scales and with higher stakes than we have seen before. Opportunities to assess learning viewed from a growing understanding of the nature and acquisition of knowledge. Opportunities to draw upon ever-expanding technological capabilities to construct scenarios, interact with examinees, capture and evaluate their performances, and model the patterns they convey. And challenges abundant, encapsulated in a single question: How can we use these new capabilities to tackle assessment problems we face today?

Long-established and well-honed assessment practices did not evolve to deal with interactive tasks, multidimensional proficiencies, and com-

plex performances. But progress is being made on many fronts, as seen, for example, in the National Board of Examiners' Primum[®] computer-based simulation tasks (Clyman, Melnick, & Clauser, 1999), Adams, Wilson, & Wang's (1997) structured multidimensional IRT models, and White and Frederiksen's (1998) guided self-evaluation in extended inquiry tasks. This work succeeds because even when it differs from familiar tests on the surface, each innovation is grounded in the same principles of evidentiary reasoning that underlie the best assessments of the past.

A vital line of current research aims to make these principles explicit, and to build from them conceptual and technological tools that can help designers put new developments to work in practice (National Research Council, 2001). Examples of work that coordinates aspects of task design, measure-

ment models, assessment delivery, and psychological research in these ways include Baker (1997, 2002), Embretson (1985, 1998), Luecht (2002), and Wilson (2005).

Our own recent work along these lines falls under the rubric of "evidence-centered" assessment design (ECD; Mislevy, Steinberg, & Almond, 2003), an approach that has been implemented variously at Educational Testing Service (Pearlman, 2001), Cisco Systems (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004), the IMS Global Learning Consortium (2000)¹, and elsewhere. We will illustrate some key ideas with examples from the Principled Assessment Design for Inquiry (PADI; Baxter & Mislevy, 2004) project, noting in particular some implications for large-scale, on-demand testing.

The next section of the paper is a brief overview of evidence-centered design. Two complementary ideas organize the effort. The first is an overarching conception of assessment as an argument from imperfect evidence. Messick (1994, p. 16) lays out the basic narrative of an assessment argument, saying that we "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?" The second idea is distinguishing layers at which activities and structures appear in the assessment enterprise,

*Robert J. Mislevy, Benjamin Building
1230-c, University of Maryland, College
Park, MD 20742; rmislevy@umd.edu.*

*Geneva D. Haertel, SRI International, 333
Ravenswood Ave., Menlo Park, CA 94025-
3493; geneva.haertel@sri.com.*

all to the end of instantiating an assessment argument in operational processes (Mislevy et al., 2003; Mislevy & Riconscente, 2006).

We then step through the layers in more detail. We see where advances in allied fields help improve the practice of assessment, and how their contributions are coordinated within and across layers. Benefits of explicitness, reusability, and common language and representations are noted throughout.

The closing discussion addresses a question from an anonymous reviewer of a proposal we submitted for a recent of the NCME meeting: Is not this all just new words for what people are already doing?

Evidence-Centered Assessment Design

Evidence-centered design views an assessment as an evidentiary argument: an argument from what we observe students say, do, or make in a few particular circumstances, to inferences about what they know, can do, or have accomplished more generally (Mislevy et al., 2003). The view of assessment as argument is a cornerstone of test validation (Kane, 1992, 2006; Messick, 1989). ECD applies this perspective proactively to test design.

Layers in the Assessment Enterprise

ECD organizes the work of assessment design and implementation in terms of layers, a metaphor drawn from architecture and software engineering (Mislevy & Riconscente, 2006). It is often useful to analyze complex systems in terms of subsystems, whose individual components are better handled at the subsystem level (Simon, 2001). Brand (1994) views buildings as dynamic objects wherein initial construction and subsequent changes take place at different timescales. He identifies six layers; from the most enduring to the most ephemeral, they are site, structure, skin, services, space plan, and stuff. Similarly, to support maintenance and troubleshooting, Cisco System's (2000) open system interconnection (OSI) reference model distinguishes seven layers of activity in computer networks: physical, data link, network, transport, session, presentation, and application. Network functions within each layer can be implemented independently and updated

without impacting the other layers. In both examples, processes and constraints interact in complex ways within layers, but cross-layer connections are more limited and tuned to the demands of the overall goal. Knowledge representations, workflow, and communications are organized in terms of layers.

Evidence-centered design applies the concept of layers to the processes of designing, implementing, and delivering an educational assessment. ECD identifies five layers. Each is characterized in terms of its role in the assessment enterprise, its key concepts and entities, and knowledge representations and tools that assist in achieving each layer's purpose. The layers are domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery (Table 1). The layers suggest a sequential design process, but cycles of iteration and refinement both within and across layers are expected and appropriate.

Domain Analysis

The *Domain Analysis* layer concerns gathering substantive information about the domain to be assessed. If the assessment being designed is to measure science inquiry at the middle school level, domain analysis would marshal information about the concepts, terminology, representational forms, and ways of interacting that professionals working in the domain use and that educators have found useful in instruction.

Examples of domain analysis can be found in the work of Webb (2006), who has described the content to be assessed in measures of achievement testing. Documents such as the National Science Education Standards (National Research Council, 1996) often provide a good starting point (state standards documents are in fact mandated foundations of accountability tests). In the area of language testing, Bachman and Palmer's (1996) taxonomy of task characteristics can be used to describe both the features of target language use (TLU) and the intended assessment. Automated methods for carrying out domain analysis, such as Shute, Torreano, and Willis's (2000) automated knowledge elicitation tool DNA (for Decompose, Network, Assess) can be tuned to provide input for Domain Modeling when a goal

such as instructional design or assessment is specified.

Transdisciplinary research on learning also tells us much about how people become proficient in domains (Ericsson, 1996) and thus what we need to assess (Mislevy, 2006). As the American Association for the Advancement of Science (1993) put it, "Some powerful ideas often used by mathematicians, scientists, and engineers are not the intellectual property of any one field or discipline. Indeed, notions of system, scale, change and constancy, and models have important applications in business and finance, education, law, government and politics, and other domains, as well as in mathematics, science, and technology. These common themes are really ways of thinking rather than theories or discoveries" (p.261). PADI's science applications revolve around paradigmatic ways of thinking, such as inquiry cycles (White & Frederiksen, 1998), knowledge representation (Greeno, 1983; Markman, 1999), model-based reasoning (Stewart & Hafner, 1994), and scaffolded learning (Brown, Collins, & Duguid, 1989).

The BioKIDS project, one of the partners in PADI, helps students learn about inquiry through increasingly independent investigations (Huber, Songer, & Lee, 2003; Songer, 2004). Consequently, the assessment tasks BioKIDS builds using PADI tools probe the degree of support that students need to, say, build scientific explanations. The first task in Figure 1, for example, provides more scaffolding for students than the second. We see in the next section how design patterns can leverage these recurring themes for building assessment tasks in different domains and at different educational levels.

As the first stage in assessment design, Domain Analysis leads us to understand the knowledge people use in a domain, the representational forms, characteristics of good work, and features of situations that evoke the use of valued knowledge, procedures, and strategies. These categories of information presage the entities and structures that appear in subsequent layers.

Domain Modeling

Using terminology from Toulmin (1958) we can say that assessment aims to make some *claim* about a student, such as proficiency in solving quadratic

Table 1. Layers of Evidence-Centered Design for Educational Assessments

Layer	Role	Key Entities	Selected Knowledge Representations
Domain Analysis	Gather substantive information about the domain of interest that has direct implications for assessment; how knowledge is constructed, acquired, used, and communicated.	Domain concepts, terminology, tools, knowledge representations, analyses, situations of use, patterns of interaction.	Representational forms and symbol systems used in domain (e.g., algebraic notation, Punnett squares, maps, computer program interfaces, content standards, concept maps).
Domain Modeling	Express assessment argument in narrative form based on information from Domain Analysis.	Knowledge, skills, and abilities; characteristic and variable task features, potential work products, potential observations.	Toulmin and Wigmore diagrams, PADI design patterns, assessment argument diagrams, “big ideas” of science.
Conceptual Assessment Framework	Express assessment argument in structures and specifications for tasks and tests, evaluation procedures, measurement models.	Student, evidence, and task models; student, observable, and task variables; rubrics; measurement models; test assembly specifications; PADI templates and task specifications.	Algebraic and graphical representations of measurement models; PADI task template; item generation models; generic rubrics; algorithms for automated scoring.
Assessment Implementation	Implement assessment, including presentation-ready tasks and calibrated measurement models.	Task materials (including all materials, tools, affordances); pilot test data to hone evaluation procedures and fit measurement models.	Coded algorithms for rendering tasks, interacting with examinees and evaluating work products; tasks as displayed; IMS/QTI representation of materials; ASCII files of item parameters.
Assessment Delivery	Coordinate interactions of students and tasks: task-and test-level scoring; reporting.	Tasks as presented; work products as created; scores as evaluated.	Renderings of materials; numerical and graphical summaries for individual and groups; IMS/QTI results files.

equations or designing experiments. The *data* are the important features of the tasks (including goals, constraints, resources, and stimulus materials) and students’ performances. The *warrant* is the reasoning that says why particular data constitute evidence for the claims. The *Domain Modeling* layer of ECD organizes information and relationships discovered in Domain Analysis along the lines of assessment arguments. Domain experts, teachers, and designers work together here to lay out what an assessment is meant to measure, and how and why it will do so, without getting tangled in the technical details that will eventually be necessary. Examples of tools for Domain Modeling are diagrams of assessment arguments (Kane, 1992), assessment argument schemas based on the “big ideas” of a given

domain (Chung, Delacruz, Dionne, & Bewley, 2003), and design patterns, an approach developed in the PADI Project that lays out, in narrative form, design options for the key elements in an assessment argument.

“Design patterns” were introduced in architecture and engineering to characterize recurring problems and approaches for dealing with them (Alexander, Ishikawa, & Silverstein, 1977; Gamma, Helm, Johnson, & Vlissides, 1994). Design patterns for assessment similarly help domain experts and assessment specialists “fill in the slots” of an assessment argument (Table 2), built around recurring themes such as the previously mentioned inquiry cycles, knowledge representation, model-based reasoning, and scaffolded performance. PADI design patterns help

assessment designers think through substantive aspects of their assessment argument, in a structure that is useful across specific domains, educational levels, and assessment purposes, and leads to the more technical work in the next layer (Mislevy et al., 2003).

Centered around some aspect of knowledge, skills, and abilities (KSAs), a design pattern suggests options for design elements that can be used to get evidence about that knowledge or skill. Table 3 shows a design pattern the BioKIDS project built, Formulating Scientific Explanations from Evidence. The two BioKIDS tasks shown earlier are both consistent with this design pattern, despite their surface differences. In the PADI design system, a design pattern appears as an online form with “slots” for each attribute. When the


Shan and Niki collected four animals from their schoolyard. They divided the animals into Group A and Group B based on their appearance as shown below:

Group A:



Group B:

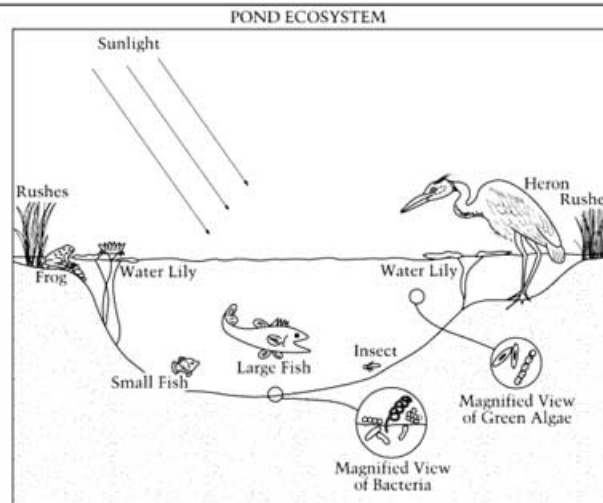


They want to place this fly  in either Group A or Group B. Where should this fly be placed?

A fly should be in Group A / Group B
(Circle one)

Name two physical characteristics that you used when you decided to place the fly in this group:

- (a)
- (b)



...If all of the small fish in the pond system died one year from a disease that killed only the small fish, what would happen to the algae in the pond? Explain why you think so.

What would happen to the large fish? Explain why you think so.

FIGURE 1. BioKIDS assessment tasks on “Formulating Scientific Explanations Using Evidence”.

design pattern is completed, it specifies elements that can be assembled into an assessment argument:

- Focal Knowledge, Skills, and Abilities (KSAs) indicate the main claim about students that tasks created from the design pattern address. In the example, it is building an explanation of observations using scientific principles. Additional KSAs may also be required to complete a task, such as whether familiarity with certain representational forms

or mathematical operations is presumed. Additional KSAs may be intentionally included in tasks, avoided, or dealt with by allowing student choice or accommodations. The example concerns building scientific explanations, but makes it clear that this can only be done using the concepts and processes of some scientific theory or model. The additional KSAs attribute makes task authors aware of design choices and their implications—including possible explanations for poor

performance due to knowledge or skills other than the targeted KSA, sources of construct-irrelevant variance in Messick’s (1989) terminology.

- Potential Work Products are things students might say, do, or make that provide information about the Focal KSAs, and Potential Observations are the aspects of the work products that constitute evidence. Potential Rubrics are ways one might evaluate work products in order to produce values of the

Table 2. Design Pattern Attributes, Definitions, and Corresponding Assessment Argument Components

Attribute	Definition	Assessment Argument Component
Rationale	The connection between the focal KSA(s) and what people do in what kinds of circumstances.	Warrant
Focal Knowledge, Skills, and Abilities	The primary knowledge/skill/abilities targeted by this design pattern.	Claim
Additional Knowledge, Skills, and Abilities	Other knowledge/skills/abilities that may be required by tasks written under this design pattern.	Claim/Alternative Explanations
Potential Work Products	Some possible things one could see students say, do, or make that would provide evidence about the KSAs.	Data
Potential Observations	Features of the things students say, do, or make that constitute the evidence.	Data about student performance.
Characteristic Features of Tasks	Aspects of assessment situations that are necessary in some form to evoke the desired evidence.	Data
Variable Features of Tasks	Aspects of assessment situations that can be varied in order to shift difficulty or focus.	Data

observations. All of these attributes concern ways of getting evidence about the targeted proficiency—and the wider the array, the better, so assessment designers can choose among a variety of ways to obtain evidence to suit the resources, constraints, and purposes of their particular situation.

- Characteristic and Variable Features of tasks specify aspects of the situation in which students act and produce work products. Characteristic Features are those that all assessment tasks motivated by the design pattern should possess in some form, because they are central to evoking evidence about the Focal KSAs. All tasks inspired by the “Formulating Scientific Explanations” design pattern, for example, involve observations and a claim about the process or pattern that explains them. Variable Features address aspects of the assessment that the assessment designer can use to affect difficulty or the focus of attention. In the Formulating Explanations example, the amount of scaffolding that a student receives is a key Variable Feature since the BioKIDS curriculum is about learning to make scientific explanations in increasingly independent situations.

Work at the domain modeling layer is important for improving the practice of assessment, especially for the higher-level reasoning and capabilities for situated actions that cognitive psychology call to our attention. Experience with experimental tasks is valuable, but it is confounded with particular domains, psychological stances, knowledge representations, and delivery vehicles. Because proficiencies are the

primary organizing category in design patterns, they help the designer keep a focus on the proficiency of interest and make sure a coherent assessment argument results. The specifics of response types, stimulus materials, measurement models, and delivery modes are then determined in light of the particular constraints and resources of the application.

The Conceptual Assessment Framework

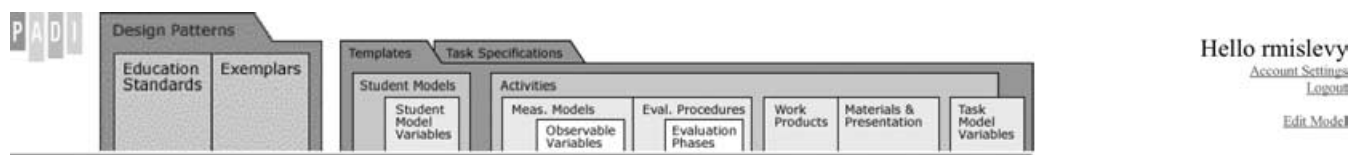
The conceptual assessment framework (CAF) concerns technical specifications for the nuts and bolts of assessments. The central models for task design are the student model, evidence models, and task models (Figure 2).² These models have their own internal logic and structures, and are linked to each other through the key elements called student-model variables, observable variables, work products, and task model variables. An assessment argument laid out in narrative form at the Domain Modeling layer is now expressed in terms of specifications for pieces of machinery such as measurement models, scoring methods, and delivery requirements. In specifying the CAF, the assessment designer makes the decisions that give shape to the actual assessment that will be generated. Details about task features, measurement models, stimulus material specifications, and the like are expressed in terms of representations and data structures that will guide their implementation and ensure their coordination.

PADI task templates are where users of the PADI design system do this work. Figure 3 shows the summary page of the task template for generating BioKIDS tasks. Some of the more detailed objects the template contains will be illustrated below.

The *Student Model* expresses what the assessment designer is trying to measure in terms of variables that reflect aspects of students’ proficiencies. Their number, character, and granularity are determined to serve the purpose of the assessment—a single student-model variable to characterize students’ overall proficiency in a domain of tasks for a certification decision, for example, or a multidimensional student model to sort out patterns of proficiency from complex performances or provide more detailed feedback. BioKIDS uses a multidimensional student model to track aspects of both content knowledge and inquiry skills such as building explanations and analyzing data (Figure 4).

A *Task Model* describes the environment in which students say, do, or make something to provide evidence. A key design decision is specifying the form in which students’ performances will be captured, i.e., the Work Product(s)—for example, a choice among alternatives, an essay, a sequence of steps in an investigation, or the locations of icons dragged into a diagram. In computer-based testing with complex tasks, reusing underlying work-product data structures streamlines authoring, implementation, and evaluation (Luecht, 2002; Scalise, 2003). The full

Table 3. “Formulating Scientific Explanations from Evidence” Design Pattern in PADI Design System



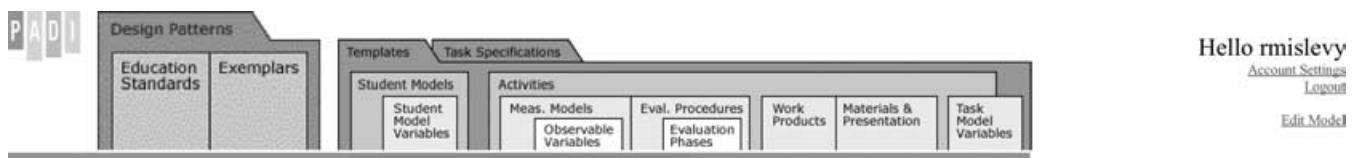
Formulating scientific explanations from evidence design pattern 1875

| View Tree | Duplicate | Export | Delete |

Title:	[Edit] Formulating scientific explanations from evidence 1	
Summary	[Edit] In this design pattern, a student develops a scientific explanation using the given evidence. The student must make a relevant claim and then justify the claim using the given evidence.	A scientific explanation consists of stating a claim and using the given data appropriately to support this claim.
Focal Knowledge, Skills, and Abilities	1 [Edit] Making a scientific explanations using evidence.	At lower levels, recognizing a scientific argument; at medium levels, producing one with structural scaffolding; at higher levels, producing one from evidence.
Rationale	1 [Edit] A key aspect of scientific inquiry is the ability to be able to propose explanations using evidence. This means forming a narrative or technical schema that connects observations in terms of underlying relationships, principles, and processes. Any particular instance requires knowledge of the underlying relationships / principles / processes.	The National Research Council lays out five essential features of classroom inquiry. Four of the five aspects involve students using evidence to create and justify explanations.
Additional Knowledge, Skills, and Abilities	1 [Edit] Conducting appropriate inquiry practices for the scientific question at hand. Formulating a logical claim based on the given data or evidence. Domain area knowledge	Necessary for the nature of the claim and rationale of the explanation
Potential Observations	1 [Edit] Weighing and sorting data/evidence The claim reflects an understanding of the data given and the requisite of scientific knowledge in terms of entities, processes, and relationships. The data that are used to support the claim is relevant, the more pieces of relevant data used, the better. The less irrelevant data used to support the claim, the better. There should be logical consistency between the evidence and the claim, in terms of the targeted domain theory / principles.	With regard to all of the potential observations, the required level of sophistication varies with the domain and level. Both primary students and medical students need to build scientific explanations, but the claims and rationales expected of medical students are vastly more technical and detailed.
Potential Work Products	1 [Edit] Matching claim and evidence (e.g., multiple choice) Spoken explanation when in a situation involving scientific concepts.	

Continued

Table 3. Continued



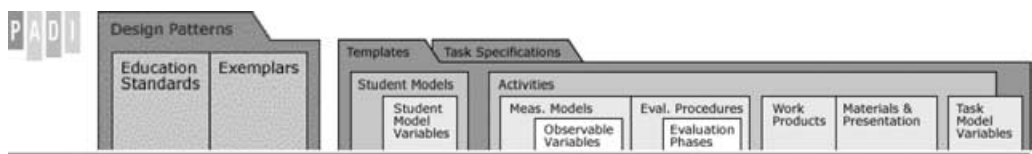
Formulating scientific explanations from evidence design pattern 1875

| View Tree | Duplicate | Export | Delete |

.		Written response – creation of claim and use of appropriate evidence to justify claim.	
Potential Rubrics	1 [Edit]	(see example tasks)	
Characteristic Features	7 [Edit]	Evidence	All items based on this design pattern have evidence. In more scaffolded questions, students may only have to choose given evidence, while in less scaffolded questions, students may have to figure out what pieces of data count as evidence.
.		Claim	All items based on this design pattern have a claim. In more scaffolded questions, the claim may be provided, while in less scaffolded questions, students may need to create a claim.
Variable Features	1 [Edit]	Level of prompting/scaffolding of creating a scientific explanation.	Less prompting makes the item more difficult for the student and thus gives better evidence about whether student is able to create scientific explanations using data on their own. More prompting makes the item easier and thus gives evidence about whether a student is able to provide an explanation using data when given the appropriate format in which to do so.
.		Level of content knowledge required	In some questions, most of the content needed to answer the question is provided by the question. However, in more difficult items, students will have to apply their content knowledge in order to answer the question.
These are parts of me	1 [Edit]	<i>Analyze data relationships.</i> A student encounters two or more sets of data organized into one or more representations, and must . . . <i>Generate explanations based on underlying scientific principles.</i> Students are asked questions about scientific phenomena that require them to . . . <i>Interpret data.</i> Students are presented with a set of data or observations and are asked to formulate an explanation . . . <i>Use data to support scientific argument.</i> A student must use data, either collected or provided, to support a scientific argument. Does the s . . .	
Educational Standards	1 [Edit]	<i>NSES 8AS11.4.</i> Develop descriptions, explanations, predictions, and models using evidence. Students should base the . . . <i>NSES 8AS11.5.</i> Think critically and logically to make the relationships between evidence and explanations. Thinking . . . <i>NSES 8AS11.6.</i> Recognize and analyze alternative explanations and predictions. Students should develop the ability . . . <i>NSES 8AS11.7.</i> Communicate scientific procedures and explanations. With practice, students should become competent . . .	

Continued

Table 3. Continued

		Hello rmislevy Account Settings Logout Edit Model
Formulating scientific explanations from evidence design pattern 1875 View Tree Duplicate Export Delete 		
. Templates	[Edit]	NSES 8AS/2.5. Scientific explanations emphasize evidence, have logically consistent arguments, and use scientific . . . Formulating Explanations, Step One Simple Template. This template represents a simple task that is well scaffolded for inquiry thinking . . . Formulating Scientific Explanations Step 1, Complex Template. This template corresponds to a task that has a high amount of scaffolding of inquiry knowledge and a . . . Formulating Scientific Explanations Step 2, moderate template. This template corresponds to a task that has a medium amount of scaffolding of inquiry knowledge and. . . BioKIDS new fish pond item. This task examines the most difficult form of question related to explanation formulation. There is. . . Formulating Scientific Explanations, Step 3 Complex. This task examines the most difficult form of question related to explanation formulation.
Exemplar Tasks	12 [Edit]	BioKIDS Step 3 complex explanations open ended question. (4) If all of the small fish in the pond system died one year from a disease that killed. . . BioKIDS step one simple explanation multiple-choice item. A biologist studying birds made the following observations about the birds. She concluded the birds . . . Scientific Explanations – Step 1, Complex Task. Biologists measured the biodiversity of animals in a city park in two different years.

Option Science System (FOSS) project, another PADI partner, designed a series of tasks that are simulations of science phenomena, to assess the science content and inquiry processes covered in the FOSS modules. Figures 5 and 6 are examples of FOSS/ASK prompts and student responses used in these items, which produce work products of a form that can be used with many tasks generated from the same template, and reused also in any number of other tasks with different content, stimulus materials, and activity patterns.

The assessment designer also specifies in a task model the forms and the key features of directives and stimulus materials, and the features of the presentation environment. For example, what resources must be available to the test taker, or what degree of scaffolding can be provided by the teacher? These decisions are guided by discussions in Domain Modeling about characteristic and variable task features. Efficiencies accrue whenever we can reuse data structures, processes, activity flows, tools, and materials; the Task Model in the CAF is where we lay out these structures and systematic, purposeful, ways for varying them.

How do we update our beliefs about a student when we observe what they say, do, or make? An *Evidence Model* bridges the Student Model and the Task Model. The two components in the evidence model—evaluation and measurement—correspond to two steps of reasoning. The *evaluation component* says how one identifies and evaluates the salient aspects of student work, in terms of values of Observable Variables. Evaluation procedures can be algorithms for automated scoring procedures, or rubrics, examples, and training materials for human scoring. Efficiencies can again be gained through reuse and modular construction, as, for example, different evaluation procedures are used to extract

different observable variables from the same work products when tasks are used for different purposes, or as different ways of implementing procedures are used to extract the same observable variables from the same work products. With specifications laid out properly, different vendors can use different algorithms to score tasks, and both human judges and automated scoring of essays produce ratings in the same form as is done with the Analytical Writing Assessment in the Graduate Management Admissions Test (Rudner, Garcia, & Welch, 2005).

Data that are generated in the evaluation component are synthesized across tasks in the *measurement model* component. Modular construction of

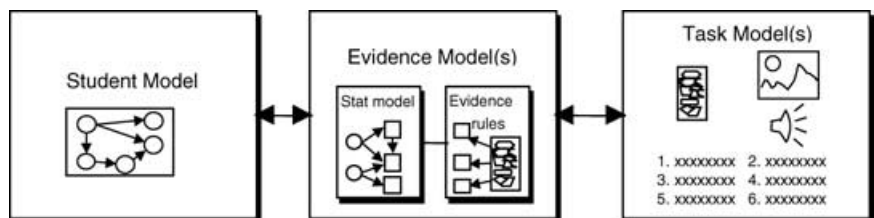


FIGURE 2. Graphic summary of the student, evidence, and task models.

BioKIDS - multidimFive | Template 1070 [View Tree | Convert to Task Spec | Duplicate | Export | Delete]

Title:	[Edit] BioKIDS - multidimFive
Summary	[Edit] This is a task specification for the entire BioKIDS test, assuming a multidimensional student model with 2 SMVs.
Type	[Edit] [View] (Modified 2004-09-25)
Student Model Summary	[Edit] Inquiry (Explanations, interpreting data, making hypotheses/predictions) + Content (Biodiversity)
Student Models	[Edit] BioKIDS 5-Dimension . Biodiversity Hypothesis Building Explanation from Evidence Reexpressing Data
Measurement Model Summary	[Edit] 16 items have MMs which vary: some are dichotomous multiple-choice models, others are bundles with both MC and open-ended models
Evaluation Procedures Summary	[Edit] Multiple choice items are dichotomous (0=incorrect; 1=correct) Open ended items are scored on a partial credit model (usually a 0-1-2 scale). Bundles are indicated where several student work products are dependent on one another.
Work Product Summary	[Edit] Some multiple choice (4-5 options) Some open-ended construction of answers to given questions
Task Model Variable Summary	[Edit]
Template-level Task Model Variables	[Edit] Amount of scaffolding . The task can guide students to think about certain concepts or can help students structure their ans... Complexity of content/materials . Amount of Data . The number of data points presented to students in graphs, tables and maps. Content area . Specific domain content under consideration Content knowledge required (simple,mod,complex) . This variable represents the amount of content knowledge needed to bring to the task in order to sol... Data Representation Format . The format of data as it is presented to students (bar graph, line graph, scatter plot, map, data ta...
Task Model Variable Settings	[Edit] [View]
Materials and Presentation Requirements	[Edit]
Template-level Materials and Presentation	[Edit]
Materials and Presentation Settings	[Edit] [View]
Activities Summary	[Edit] One activity per item because, for a bundled item, the activity helps associate the MM with the proper Eval Procedure in a way that the Gradebook can discern.
Activities	[Edit] BioKIDS pre/posttest activity multidimFive (all MMs) .
Tools for Examinee	[Edit] Paper and pencil/pen This test is entirely written

FIGURE 3. A BioKIDS template within PADI design system.

measurement models assembles pieces of IRT or other models. One development of recent interest is assembling tasks and corresponding measurement models in accordance with task model variables (Embretson, 1998). Much

can be gained especially when evidentiary relationships in complex tasks and multivariate student models are expressed in reusable measurement model fragments. BioKIDS handles the conditional dependencies in their

Claim and Explanation tasks (Gotwals & Songer, 2006) in a modular way, using the same “bundled” structure to model structurally similar responses from the many tasks that can be generated from their task templates. Using the BioKIDS

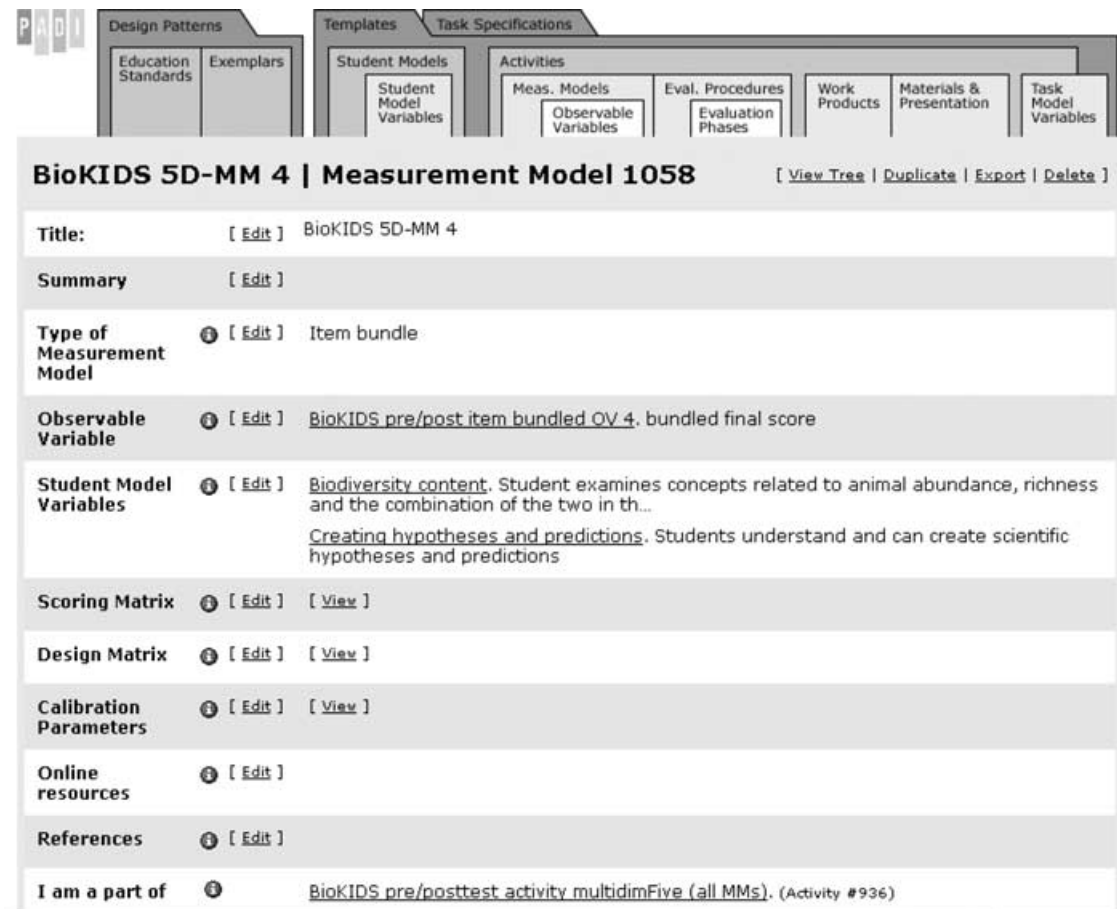


FIGURE 4. A BioKIDS measurement model within the PADI design system, as viewed through the user interface.

PADI templates, task authors create unique complex tasks but know ahead of time “how to score them.”

Figure 7 is a high-level representation of a PADI template as a unified modeling language (UML) diagram, the data structure behind what a designer sees. Software engineers use this representation to create authoring systems, databases, and delivery systems that work with each other using the design objects. When the designer works with the interface, XML files are produced in this form, a representation used to share and transport information, and to structure input and output to automated procedures for building or delivering tasks.

Assessment designers do not work directly with UML representations, and they generally should not even have to work with the generic design system. They should be able to use interfaces that are organized around their jobs, in their language, with the information they need. For this reason, PADI has developed a collection series of “wizards” (and a tool for creating them; Hamel & Schank, 2006) that guide de-

signers through various more particular jobs—for example, building a template, creating a specification for a new task based on an existing template, and selecting activities from a multi-stage investigation in a way that meets targeted time constraints, provides summary or diagnostic feedback, and focuses on designated proficiencies. Figure 8 is one of the screens from the wizard for designing a new FOSS/ASK tasks. Each screen in a series of four asks questions that help a designer select or create elements for building a new task from the FOSS/ASK template. Wizards hide the complexity of the underlying structure and maintain coherence among the student, task, and evidence models (see Wickham, Mayhew, Stoll, Toley, & Rouiller, 2002, on designing wizards). In this way they help users who are not experts in assessment design or psychometrics use the design system.

There are several considerable advantages to explicating the objects in this design layer. Constructing coordinated forms helps organize the work of the different specialists who are in-

involved in designing complex assessments. Because the CAF models are themselves nearly independent, they are readily recombined when the same kinds of tasks are adapted for other purposes—from summative to formative uses, for example, by using a finer-grained student model. Common data structures encourage the development of supported or automated processes for task creation (e.g., Irvine & Kyllonen, 2002), evaluating work products (e.g., Williamson, Mislevy, & Bejar, 2006), and assembling measurement models (e.g., Rupp, 2002; von Davier, 2005). These features are especially important for computer-based tasks that are costly to author and implement, such as interactive simulations (see, for example, Niemi & Baker, 2005, on task design; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002, on measurement models; Luecht, 2002, on authoring and assembly; and Stevens & Casillas, 2006, on automated scoring). Bringing down the costs of such tasks requires exploiting every opportunity to reuse arguments, structures, processes, and materials.

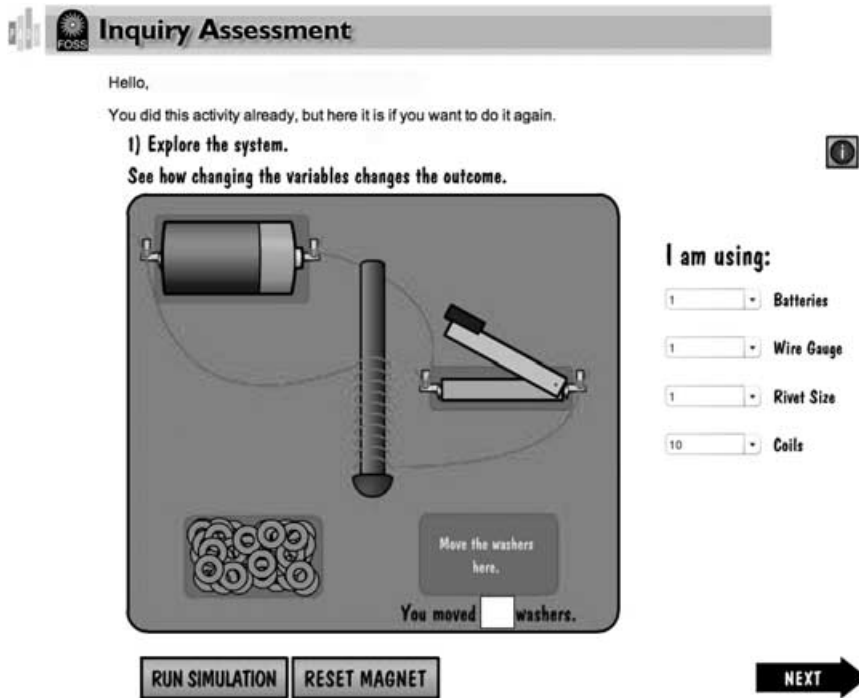


FIGURE 5. Prompt from FOSS/ASK simulation.

Assessment Implementation

The *Assessment Implementation* layer of ECD is about constructing and preparing all of the operational elements specified in the CAF. This includes authoring tasks, finalizing rubrics or automated scoring rules, estimating the parameters in measurement models, and producing fixed test forms or algorithms for assembling tailored tests. All of these activities are familiar in current tests and are often quite efficient in and of themselves. The ECD approach links the rationales for each back to the assessment argument, and provides structures that offer opportunities for reuse and interoperability. Compatible data structures leverage the value of systems for authoring or generating tasks, calibrating items, presenting materials, and interacting with examinees (e.g., Baker, 2002; Niemi, 2005; and Vendlinsky, Niemi, & Baker, in press). To this end, PADI data structures are compatible with the IMS's QTI (Question and Test Interoperability) standards for computer-based testing data and processes.

PADI tools for the implementation layer include a calibration engine for the models in the Multidimensional Random Coefficients Multinomial Logit Model family (MRCMLM; Adams, Wilson, & Wang, 1997) and some of the PADI wizards help test developers cre-

ate inquiry tasks from PADI templates from the FOSS and GLOBE examples.

Another PADI demonstration is the Mystery Powders simulation-based investigation tasks (Siebert, Hamel, Haynie, Mislavy, & Bao, 2006). Building on a classical performance assessment in which students use tests to determine which of several powders their "mystery powder" consists of, this

demonstration illustrates all layers of the ECD framework. We will say more about Mystery Powders in the next section, but in regard to the implementation layer, Mystery Powders constructs tasks and interacts with students on the fly using a prespecified library of stimulus materials (including video clips of reactions of powders to tests) and specifications from a PADI template. Both the Mystery Powders tasks and the FOSS/ASK tasks illustrate how interactive computer-based inquiry tasks can be implemented on a large scale, such as required for No Child Left Behind testing, to obtain evidence about aspects of using science that are difficult to assess in traditional formats. This possibility is particularly attractive as several states are now migrating to computer-based test administration.

Assessment Delivery

The *Assessment Delivery* layer is where students interact with tasks, their performances are evaluated, and feedback and reports are produced. The PADI project uses the four-process delivery system described in Almond, Steinberg, and Mislavy (2002), which is also the conceptual model underlying the IMS/QTI standards. This way of parsing assessment systems can be used to describe not only computer-based testing procedures, but also paper-and-pencil tests, informal classroom tests, or tutoring systems. Common

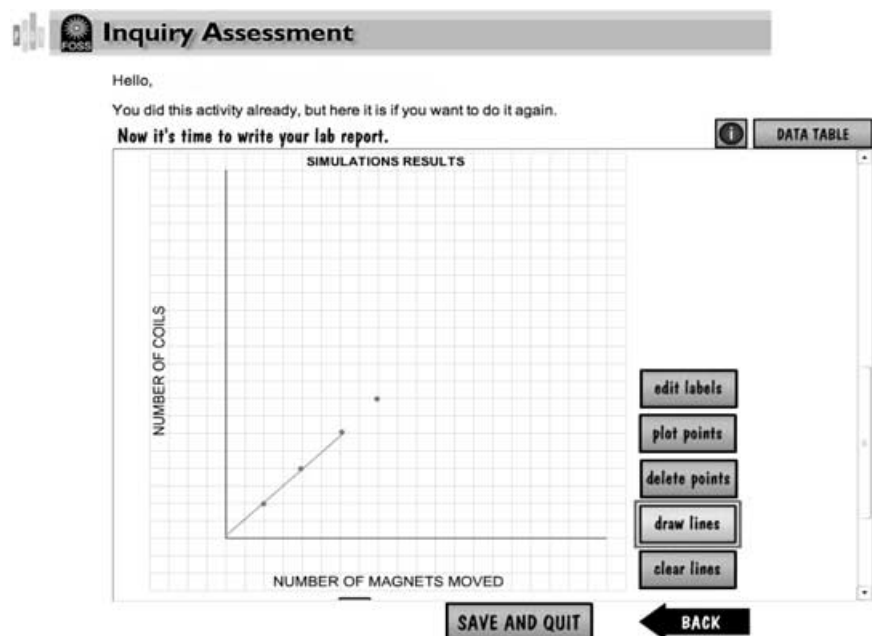


FIGURE 6. Student work product from FOSS/ASK simulation.

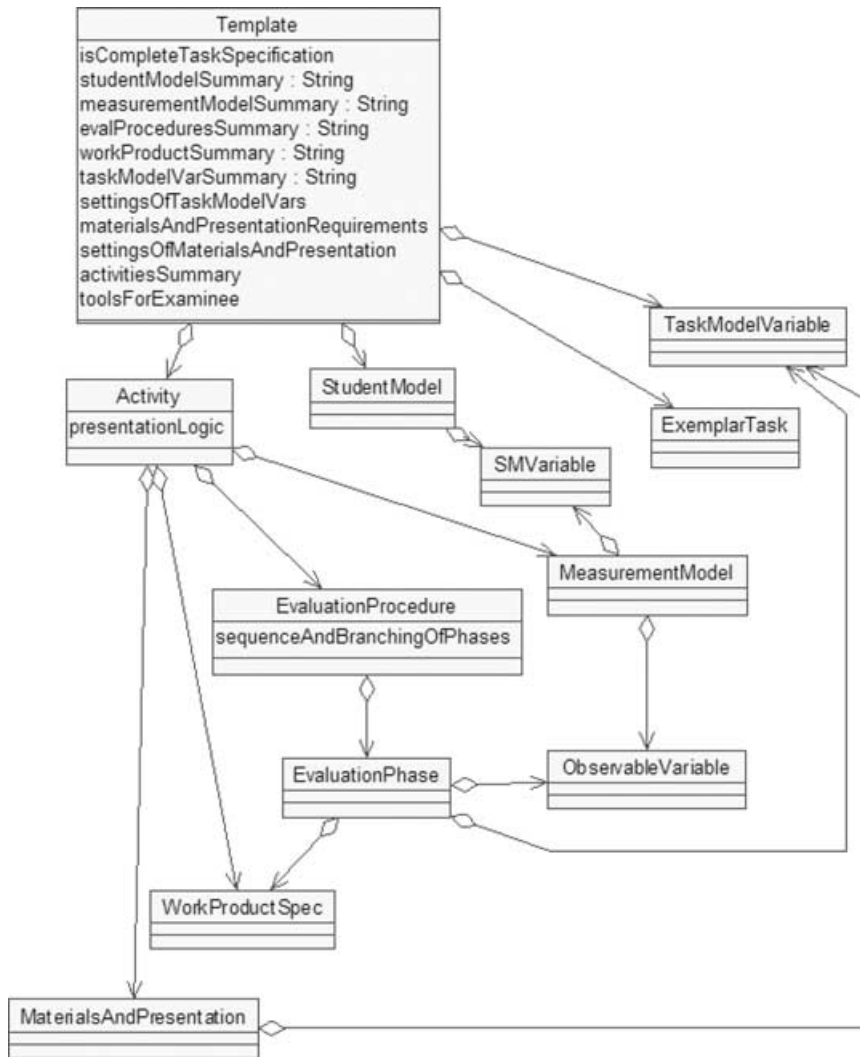


FIGURE 7. High-level UML representation of the PADI object model.

language, common data structures, and a common partitioning of activities all promote reuse of objects and processes, and interoperability across projects and programs. When an assessment is operating, the processes pass messages among one another in a pattern determined by the test's purpose. All of the messages are either data objects specified in the CAF (e.g., parameters, stimulus materials) or are produced by the student or other processes in data structures that are specified in the CAF (e.g., work products, values of observable variables).

Assessment operation is represented as four principal processes. The *activity selection process* selects a task or activity from the task library, or creates one in accordance with templates in light of what is known about the student or the situation. The *presentation process* is responsible for presenting

the task to the student, managing the interaction, and capturing work products. Work Products are then passed to the *evidence identification process*, or task-level scoring. It evaluates work using the methods specified in the Evidence Model. It sends values of Observable Variables to the *evidence accumulation process*, or test-level scoring, which uses the Measurement Models to summarize evidence about the student model variables and produce score reports. In adaptive tests this process provides information to the *activity selection process* to help determine what tasks to present next.

As with Assessment Implementation, many assessment delivery systems exist and many are quite efficient in the settings for which they were developed. Reusability and interoperability are the watchwords here, particularly for web- and computer-based testing. The ECD

framework helps designers develop assessment materials and processes that fit current standards and, more generally, accord with the overarching principles. Such efforts help bring down the costs of developing, delivering, and scoring innovative assessments at the large scale required in large-scale testing.

The PADI Mystery Powders demonstration mentioned previously illustrates the deep interconnections among narrative, substantively grounded assessment arguments, specifications for the technical details of the operational elements of assessments, and the processes and operations of assessment delivery, all in terms of the structures of the PADI framework. This example is of particular interest to assessment designers and measurement specialists confronting the challenges of large-scale assessments. Large-scale assessment designers are charged with measuring challenging content with great precision and doing it under constraints of time and dollars—such conditions suggest that the advantages conferred through the use of technology may provide an avenue for accomplishing these goals. The Mystery Powders prototype is an exemplar for developing other technology-based assessment systems based on the four-process architecture, in particular with specifications in terms of the PADI framework and with messaging consistent with IMS/QTI standards.

Is Not This Just New Words for Things We Already Do?

So, what is the bottom line: Is evidence-centered design just a bunch of new words for things we are already doing? There is a case to be made that it is. All of the innovations sketched above—in cognitive psychology, learning in domains, measurement models, task design, scoring methods, web-based delivery, and more—have been developed by thousands of researchers across many fields of study, without particular regard for ECD. So too have new assessment types arisen, each in their stead. And established and efficient procedures for familiar assessments have been evolving for decades, continually improving in increments. Changing vocabularies and representational forms would like as not slow them down, as long as their current

Activities

Please select the Activity you wish to use.

Please select exactly 1 entry.

- ASK Model Scenario Activity Blueprint** Students read about an experiment and the findings, then answer questions about it. Based on a "model" experimental process. This is a blueprint to be filled out for an actual activity.
- ASK Exception Scenario Activity Blueprint** Students read about an experiment and the findings, then answer questions about it. These are not "model" approaches.
- ASK Performance Activity Blueprint** Students design and conduct an experiment and interpret the data to draw conclusions. Student is presented with collection of materials (some irrelevant) and measuring tools (some irrelevant). Students are assessed on selecting appropriate materials, measuring tools; controlling variables; drawing appropriate conclusions; and providing relevant and sufficient evidence. Activities may be delivered by computer or paper and pencil. Work products will always be paper and pencil.

back next

FIGURE 8. A screen from the FOSS/ASK task completion wizard.

goals and processes suit their aims and resources.

But efficiency just within assessments can impede efficiency across assessments. Efficient work within large-scale assessments takes place because each contributor knows his or her job, but connections among the work they do remain implicit. Modifying an assessment is difficult, since what appear to be improvements from one vantage point conflict with other parts of the system in unforeseen ways. Elements or processes that could in principle be shared across assessments are not, because their data structures are incompatible or delivery stages are collapsed differently. Analyzing existing assessments in terms of common vocabularies and representational forms across the ECD layers helps bring out the fundamental similarities across assessments that can look very different on the surface, and alert us to opportunities for improvement.

Even greater gains accrue for new kinds of tests, both conceptually and technically. The conceptual advantages come from grounding the design process from the beginning on the assessment argument, in the form of tools like design patterns. Thinking through how to assess new or complex proficiencies, as in science inquiry and task-based

language assessment, is best done at a layer that focuses on the conceptual argument, capable of being implemented in different ways rather than being entangled with implementation or delivery choices. This work is a natural bridge between conceptual developments reflected in research and standards on the one hand, and practical testing methods on the other. Work at this layer improves practice in its own ways for large-scale, classroom, certification, or other testing venues.

The technical advantages come about because no existing process can be pulled off the shelf and implemented in toto. More original design work is therefore necessary to rationalize, implement, and deliver a new kind of, say, simulation tasks. ECD's language, representational forms, and a unified perspective of the assessment enterprise guide planning and coordinate work in developing tasks and operational methods. They entail laying out the assessment argument, clarifying design choices, and coordinating the development of operational elements. They encourage at every step along the way the recognition and exploitation of efficiencies from reuse and compatibility. Moreover, they provide a principled framework to work through accommodation and universal design, at

the level of the validity argument as well as delivery issues. Hansen, Mislevy, Steinberg, Lee, and Forer (2005), for example, provide an algorithm for presenting materials during assessment delivery in a manner that circumvents construct-irrelevant access or response requirements a given student would have encountered faced with standard test forms.

Evidence-centered design is a framework that does indeed provide new words for things we are already doing. But it helps us understand what we are doing at a more fundamental level. And it sets the stage for doing what we do now more efficiently, and learning more quickly how to assess in ways that we do not do now, either because we do not know how, or cannot afford to.

Notes

¹This organization was originally the Instructional Management Systems project. The term raised more questions than it answered, so the name was changed to the IMS Global Learning Consortium.

²Defined abstractly in Mislevy, Steinberg, and Almond (2003), they can be implemented in different specific forms, as in Wilson's (2005) four-model parsing of the system and PADI template objects that catenate evidence and task models.

Acknowledgments

This material is based on work supported by the National Science Foundation under grant REC-0129331 (PADI Implementation Grant). Any opinions, findings, and conclusions or recommendations expressed in this materials are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are grateful to the editor for his helpful comments.

References

- Adams, R., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.
- 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>.
- American Association for the Advancement of Science (AAAS) (1993). *Benchmarks for scientific literacy*. New York: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Baker, E. L. (1997, Autumn). Model-based performance assessment. *Theory Into Practice*, 36, 247–254.
- Baker, E. L. (2002). Design of automated authoring systems for tests. Proceedings of technology and assessment: Thinking ahead proceedings from a workshop (pp. 79–89). Washington, DC: National Research Council, Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education.
- Baxter, G., & Mislevy, R. J. (2004). *The case for an integrated design framework for assessing science inquiry* (CSE Technical Report 638). Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESSST), Center for Studies in Education, UCLA.
- Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *The International Journal of Testing*, 4, 295–301.
- Brand, S. (1994). *How buildings learn: What happens after they're built*. New York: Viking-Penguin.
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18(1), 32–42.
- Chung, G. K., Delacruz, W. K., Dionne, G. B., & Bewley, W. L. (2003, December). Linking assessment and instruction using ontologies. Proceedings of the I/ITSEC, Orlando, FL.
- Cisco Systems (2000). *Internetworking technology basics* (3rd ed.). Indianapolis, IN: Cisco Systems.
- Clyman, S. G., Melnick, D. E., & Clauser, B. E. (1999). Computer-based case simulations from medicine: Assessing skills in patient management. In A. Tekian, C. H. McGuire, W. C. McGahie (Eds.), *Innovative simulations for assessing professional competence* (pp. 29–41). Chicago: University of Illinois, Department of Medical Education.
- Embretson, S. E. (1985). A general latent trait model for response processes. *Psychometrika*, 49, 175–186.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 380–396.
- Ericsson, K. A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performances, sports, and games*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns*. Reading, MA: Addison-Wesley.
- Gotwals, A. W., & Songer, N. B. (2006). *Cognitive predictions: BioKIDS implementation of the PADI assessment system* (PADI Technical Report 10). Menlo Park, CA: SRI International.
- Greeno, J. G. (1983). Conceptual entities. In D. Gentner & A. L. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hamel, L., & Schank, P. (2006). *A Wizard for PADI assessment design* (PADI Technical Report 11). Menlo Park, CA: SRI International.
- Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests within a validity framework. *System: An International Journal of Educational Technology and Applied Linguistics*, 33, 107–133.
- Huber, A. E., Songer, N. B., & Lee, S.-Y. (2003). *A curricular approach to teaching biodiversity through inquiry in technology-rich environments*. Paper presented at the annual meeting of the National Association of Research in Science Teaching (NARST), Philadelphia.
- IMS Global Learning Consortium (2000). *IMS question and test interoperability specification: A review* (White Paper IMSWP-1 Version A). Burlington, MA: IMS Global Learning Consortium.
- Irvine, S. H., & Kyllonen, P. C. (Eds.) (2002). *Item generation for test development*. Hillsdale, NJ: Erlbaum.
- Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527–535.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.
- Luecht, R. M. (2002). From design to delivery: Engineering the mass production of complex performance assessments. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Markman, A. B. (1999). *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. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Westport, CT: American Council on Education/Praeger.
- Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–67.
- Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education*, 15, 363–378.
- National Research Council (1996). *National Science Education Standards*. Washington, DC: National Academy Press.
- National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment, J. Pellegrino, R. Glaser, & N. Chudowsky (Eds.). Washington, DC: National Academy Press.
- Niemi, D. (2005, April). Assessment objects for domain-independent and domain specific assessment. In F. C. Sloane & J. W. Pellegrino (co-Chairs), *Moving technology up-design requirements for valid, effective classroom and large-scale assessment*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.
- Niemi, D., & Baker, E. L. (2005, April). Reconciling assessment shortfalls: System requirements needed to produce learning. In F. C. Sloane & J. W. Pellegrino (co-Chairs), *Moving technology up-design requirements for valid, effective classroom and large-scale assessment*. Presentation at the annual meeting of the American Educational Research Association, Montreal.
- Pearlman, M. (2001). Performance assessments for adult education: How to design performance tasks. Presented at the workshop “Performance Assessments for Adult Education: Exploring the Measurement Issues,” hosted by the National Research Council’s Board on Testing and Assessment, December 12–13, Washington, DC.
- Rudner, L., Garcia, V., & Welch, C. (2005). An evaluation of Intellimetric™ essay scoring system using responses to GMAT® AWA prompts (GMAC Research report number RR-05-08). McLean, VA: Graduate Management Admissions Council.
- Rupp, A. A. (2002). Feature selection for choosing and assembling measurement models: A building-block-based organization. *International Journal of Testing*, 2, 311–360.
- Scalise, K. (2003). *Innovative item types and outcome spaces in computer-adaptive assessment: A literature survey*. Berkeley Evaluation and Assessment Research (BEAR) Center, University of California at Berkeley.
- Shute, V. J., Torreano, L., & Willis, R. (2000). DNA: Towards an automated knowledge elicitation and organization tool. In S. P.

- Lajoie (Ed.), *Computers as Cognitive Tools*, Volume 2 (pp. 309–335). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Siebert, G., Hamel, L., Haynie, K., Mislevy, R., & Bao, H. (2006). *Mystery powders: An application of the PADI design system using the four-process delivery system*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Simon, H. A. (2001). *The sciences of the artificial* (4th ed.). Cambridge, MA: MIT Press.
- Songer, N. B. (2004) Evidence of complex reasoning in technology and science: Notes from Inner City Detroit, Michigan, USA. IPSI-2004 Pescara Conference, Italy.
- Stevens, R., & Casillas, A. (2006). Artificial neural networks. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer based testing* (pp. 259–312). Mahwah, NJ: Erlbaum Associates.
- Stewart, J., & Hafner, R. (1994). Research on problem solving: Genetics. In D. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 284–300). New York: Macmillan.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge, England: Cambridge University Press.
- Vendlinsky, T. P., Niemi, D., & Baker, E. L. (in press). Objects and templates in authoring problem-solving assessments. In E. L. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations*. Mahwah, NJ: Erlbaum.
- von Davier, M. (2005). A class of models for cognitive diagnosis. *Research Report RR-05-17*. Princeton, NJ: ETS.
- Webb, N. (2006). Identifying content for student achievement tests. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 155–180). Mahwah, NJ: Erlbaum.
- White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, 16, 3–118.
- Wickham, D., Mayhew, D., Stoll, T., Toley, K., & Rouiller, S. (2002). *Designing effective wizards: A multidisciplinary approach*. Upper Saddle River, NJ: Prentice Hall.
- Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.) (2006). *Automated scoring of complex tasks in computer based testing*. Mahwah, NJ: Erlbaum Associates.
- Wilson, M. R. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.