Leveraging Evidence-Centered Assessment Design in Large-Scale and Formative Assessment Practices

Prepared for:

2010 Annual Meeting of the National Council on Measurement in Education (NCME)

April 29th – May 4th, Denver, CO

Prepared by:

Eric Snow, SRI International
Geneva Haertel, SRI International
Dennis Fulkerson, Pearson
Mingyu Feng, SRI International
Paul Nichols, Pearson

Acknowledgments

This material is based on work supported by the National Science Foundation under grant DRL-0733172 (ECD Large-scale Grant).

Disclaimer

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

Abstract

The last two decades have seen innovations in the world of educational assessment that require a different language and novel representations to describe their affordances. Promising research programs are developing conceptual and technological tools to help assessment developers leverage these innovations in practice. Large-scale assessment development practices have not yet taken full advantage of these innovations and associated research programs. This is due in part to the demands of operational programs, the cost and logistical constraints of large-scale testing, and the challenges of integrating innovations with established practices. Evidence-centered assessment design (ECD) is an integrated approach to constructing educational assessments in terms of evidentiary arguments that can be leveraged by large-scale and formative assessment developers to improve validity and maximizing efficiencies in the design, development and delivery processes. This paper depicts ECD as a series of integrated layers describing an assessment design process that includes analyzing and modeling domains, specifying arguments in terms of student, task and evidence models, and implementing the assessment and executing operational processes. Current findings from *An Application of Evidence-Centered Design (ECD) to a State's Large Scale Science Assessment,* a 5-year project funded by the National Science Foundation, are used to highlight principles and structures of ECD – standards alignment, narrative structures, design patterns and task templates – that were identified as ways to leverage the *Minnesota Comprehensive Assessment-II, Science Assessment (MCA-II)* assessment design, development, and delivery processes. The presentation of the ECD principles and structures, particularly design patterns, will be periodically extended beyond its current use in leveraging large-scale assessment design to highlight how it might be leveraged in formative assessment design, as well.

Introduction: Leverage Points in Large-Scale Test Development

The last two decades have seen innovations in the world of educational assessment - the emergence of interactive assessment tasks, multidimensional proficiencies, and the modeling of complex performances and cognitive processes - that require a different language and novel representations to describe their affordances. Over the past decade several promising research programs (e.g., Leighton & Gierl, 2007; Gorin, 2006; Wilson, 2005; Mislevy, Steinberg, & Almond, 2003; Baker, 2002; Luecht, 2002; Embretson, 1998) have been implemented that aim to develop conceptual and technological tools that can help assessment designers and developers leverage the new innovations in practice (NRC, 2001).

Long-established and well-honed practices for developing large-scale assessments, however, have not always taken full advantage of these innovations and associated research programs. This is due in part to the demands of operational programs, the cost and logistical constraints of large-scale testing, and the challenges of integrating innovations with established practices. Furthermore, it is not always clear whether and how opportunities and processes, or "leverage points" (Mislevy, Steinberg, Almond, Haertel, & Penuel, 2003), can be employed to enhance large-scale assessment design, development and delivery processes. The issue is further complicated because the leverage points that exist in large-scale assessment design, development and delivery processes necessarily vary based on test development process, as well as the focus and relative strengths (i.e., "levers") of the research program being applied.

The development and application of the evidence-centered assessment design (ECD) framework (Mislevy, Steinberg, & Almond, 2003) is a research program that aims to make the work of assessment design and development, particularly for large-scale tests, more efficient and potentially more valid than current practices. ECD has been applied variously at Educational Testing Service (ETS; Pearlman, 2001), Cisco Systems (Behrens, Mislevy, Bauer, Williamson, and Levy, 2004), and the IMS Global Learning Consortium (2000). Additionally, ECD was used as the foundation for the Principled Assessment Design for Inquiry project (PADI; Mislevy & Haertel, 2006; Baxter and Mislevy, 2004) and, more recently, to guide the development and revision of Advanced Placement Exams at the College Board (Huff & Plake, 2009).

This paper extends earlier work that focused on "leveraging" evidence-centered design (ECD) in large-scale test design, development and delivery processes (e.g., Haertel & Mislevy, 2008; Mislevy 2007; Mislevy & Haertel, 2006; Mislevy, Steinberg, Almond, Haertel, & Penuel, 2003). The second section reviews the five layers of ECD that test designers can use for explicating the measurement domain and relevant constructs; structuring assessment elements such as tasks, rubrics, and psychometric models into an assessment argument; and for implementing these elements in an operational assessment. The third section introduces current work applying the ECD layers to the design and development of a large-scale assessment – the *Minnesota Comprehensive Assessment-II, Science Assessment (MCA-II)*, and provides a high level overview of the primary stages of the *MCA-II* development process. The fourth section highlights features of ECD – standards alignment, narrative structures, design patterns, and task

templates – that were identified as ways to "leverage" the *MCA-II* assessment design, development, and delivery processes. The closing discussion will highlight additional possibilities for leveraging ECD in the large-scale assessment development process and offer some preliminary thoughts on possible leverage points in the classroom assessment development process.

The presentation will closely follow the organization of the paper by: 1) providing a high level overview of the five layers of ECD; 2) describing the current project applying ECD layers to the design and development of the *MCA-II*, including an overview of the primary stages of the *MCA-II* development process; 3) describing how features of ECD, particularly design patterns, were used to improve the efficiency and validity of the *MCA-II* design, development and delivery processes; and 4) highlighting additional possibilities for leveraging features of ECD in large-scale and classroom assessment development processes.

Layers of Evidence-Centered Design

This section provides a high-level review of the layers of the evidence-centered design (ECD) framework. The five layers shown in Table 1, adapted from Mislevy and Haertel (2006), focus in turn on the substantive domain (Layer 1); the assessment argument (Layer 2); the structure of assessment elements such as tasks, rubrics, and psychometric models (Layer 3); the implementation of these elements (Layer 4); and the way they function in an operational assessment (Layer 5).

Table 1: Layers of ECD

Layer	Role
Layer 1: Domain Analysis	Gather substantive information about the domain of interest that has direct implications for assessment; how knowledge is constructed, acquired, used, and communicated
Layer 2: Domain Modeling	Express assessment argument in narrative form based on information from Domain Analysis
Layer 3: Conceptual Assessment Framework	Express assessment argument in structures and specifications for tasks and tests, evaluation procedures, and measurement models
Layer 4: Assessment Implementation	Implement assessment, including presentation-ready tasks and calibrated measurement models
Layer 5: Assessment Delivery	Coordinate interactions of students and tasks, task-and test-level scoring, reporting

Layer 1: Domain Analysis

The *Domain Analysis* layer of ECD is concerned with gathering substantive information about the domain of interest that holds meaning for assessment. This includes the content,

concepts, terminology, and representational forms that people working in the domain use. It also includes conceptions about the nature of knowledge in the targeted domain as well as the way people use the knowledge to solve problems. The content, concepts and representational forms that make up the heart of the domain analysis layer of ECD can be situated in national and state academic content standards, as well as in classroom curricula and practices. Work that takes place in domain analysis provides the grist for an assessment argument that is further specified in the entities and structures that appear in subsequent ECD layers.

Layer 2: Domain Modeling

In the *Domain Modeling* layer of ECD, information and relationships discovered in domain analysis are organized in a narrative form that serves as a high-level introduction to the assessment argument that will support the new assessment being designed. The work in this layer is a transition from specialized knowledge about the domain to the specialized knowledge about the more technical machinery of assessment, which takes place in the third layer. Toulmin diagrams (1958) are examples of tools for organizing assessment arguments at a narrative level (e.g., Kane, 1992), as are design patterns. As discussed in this section, design patterns are meant to guide the design of families of assessment tasks organized around aspects of proficiency, which could be implemented in many different ways depending on the particulars of the testing contexts (i.e., statewide/summative, or classroom/formative).

Although each assessment application is to some degree unique in its contents, purposes, and contexts, there are certain principles and relationships that all will share simply because all are assessments. For this reason one may gain advantage by embedding these principles in processes and knowledge representations. Architect Christopher Alexander (1977) coined the term *design pattern* in the mid-1970s. A *design pattern* is the core of a solution to a problem that occurs repeatedly in our environment — but at a level of generality that the approach can be applied in many situations while adapting to the particulars of each case. The same idea was adapted by software engineers to help designers tackle programming problems that recur in different guises (Gamma et al., 1994). For these engineers, *design patterns* provide structured insights into conceptual problems and solutions above the level of specific programming languages and implementation environments.

Analogous forms called assessment *design patterns* were developed by Mislevy et al. (2003) to support the design of tasks for assessing science inquiry in the Principled Assessment Designs for Inquiry (PADI) project. Like designing tests of communicative competence, designing science inquiry tasks is a challenge to standard assessment development practice (i.e., inquiry is regarded in the assessment community as a construct that is hard to assess). It calls for extended performances, cycles of hypothesis and testing, and, often, technologies such as automated scoring and computer-based simulation environments. Design patterns provide assessment designers with an high-level approach to tackle challenging issues by scaffolding the thinking that must proceed the particular technical decisions required in the development of the actual tasks, identification of psychometric models, and articulation of decision rules required for scoring tasks. Assessment *design patterns* organize information about the targeted proficiencies,

performance, and use situations in terms of the structure of assessment arguments. They serve as an in-between layer that connects the content of an assessment argument to the structure of the argument.

In particular, each *design pattern* builds around the general form of an assessment argument, concerning the knowledge or skill one wants to address (examples in science inquiry include model-based reasoning and designing experiments), the kinds of observations that can provide evidence about acquisition of this knowledge or skill, and the features of task situations that allow students to provide this evidence. Explicating the assessment structure in a narrative form with slots to be filled, *design patterns* arrange an underlying assessment argument into attributes that can subsequently be instantiated in particular operational tasks. Because the structure of a *design pattern* implicitly contains the structure of an argument in general, and an assessment argument in particular, filling in the design pattern slots simultaneously renders explicit the relationships among the pieces of the design pattern attributes in terms of the roles they play in argumentation based on Toulmin's diagram, as well as the roles they play in an assessment argument based on Messick's components (see Table 2, adapted from Mislevy and Haertel, 2006).

Work at the domain modeling layer is important for improving the practice of assessment, especially for the valid measurement of higher- level reasoning and capabilities for situated actions that cognitive psychology call to our attention (e.g., scientific inquiry). Experience with experimental tasks is valuable, but it is confounded with particular domains, psychological stances, knowledge representations, and delivery vehicles. Because constructs are the primary organizing category in design patterns, they help the designer keep a focus on the construct 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.

Table 2: Design Pattern Attributes, Definitions & Corresponding Messick and Toulmin Argument Components

Design Pattern Attribute	Attribute Definition	Messick Assessment Argument Component	Toulmin Assessment Argument Component
Rationale	The connection between the focal KSAs and what people do in what kinds of circumstances.	**Student Model/Claim** What construct (complex of student attributes) should be assessed?	Warrant
Focal Knowledge, Skills & Abilities	The primary KSAs targeted by the Design Pattern.		Claim
Additional Knowledge, Skills & Abilities	Other KSAs that may be required by tasks written using 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.	**Evidence Model/Actions** What behaviors should reveal the construct?	Data about student performance
Potential Observations	Features of the things students say, do, or make that constitute the evidence.		Data about student performance
Characteristic Task Features	Aspects of assessment situations that are necessary in some form to elicit desired evidence.	**Task Model/Situation** What tasks should elicit those behaviors?	Data about assessment context/situation
Variable Task Features	Aspects of assessment situations that can be varied in order to shift difficulty or focus.		Data about assessment context/situation

Layer 3: Conceptual Assessment Framework

The elements and processes that are needed to implement an assessment that embodies the argument are specified in the *Conceptual Assessment Framework* (CAF). In the CAF, structures such as task templates, task specifications, and scoring algorithms give concrete shape to the assessments a developer needs to create. These decisions are specific and detailed, and must reflect the purposes, constraints, and resources of the intended use. Work in the CAF layer converts the assessment arguments sketched in domain modeling into operational terms. Whereas a design pattern provides a high level focus that organizes issues that need to be considered whenever one assesses a targeted aspect of proficiency, task templates and test specifications provide a detailed blueprint for designing and writing tasks with specified properties that suit the purposes, constraints, and resources of the particular testing context. Tasks created from the same specification are equivalent in terms of purpose, cost, and so on, and can be used interchangeably.

The work in the CAF layer is organized around specifying three primary models: student, evidence and task models (See Messick, 1994; Almond, Steinberg, & Mislevy, 2002). Each of these models has their own internal logic and structures, and is linked to other models through key elements called student-model variables, observable variables, work products, and task model variables.

Student Model. A Student Model expresses what the assessment designer is trying to measure in terms of variables that reflect aspects of students' proficiencies. The number of student model variables identified, as well as their character and granularity are determined by the purpose of the assessment — a single student-model variable to characterize students' overall proficiency in the domain of tasks for a certification decision, for example, or a multidimensional student model to sort out patterns of proficiency from complex performances or to provide more detailed feedback.

Task Model. A Task Model describes the environment in which students say, do, or make something to provide evidence for their proficiencies. 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 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 the Domain Modeling layer 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. A critical question

remains: How do we update our beliefs about a student when we observe what they say, do, or make?

Evidence Model. 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 venders 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 multiple tasks in the measurement model component and are used to inform the revision and development of student model variables. IRT, latent class models, Bayes nets, and other types of measurement models are applied to the student scores that are the result of the application of the evaluative component. In the past decade there has been increasing interest in 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.

There are several considerable advantages to explicating the model components in the CAF design layer. Constructing coordinated forms helps organize the work of the different specialists who are 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 finer-grained student and evidence models. 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 performance-based and computer-based tasks that are costly to author and implement, such as interactive simulations (e.g., see Niemi & Baker, 2005, Quellmalz, Timms, & Schneider, 2009, 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.

Layer 4: 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, 2008). In order to maximize this leverage point, PADI data structures are compatible with the IMS's QTI (Question and Test Interoperability) standards for computer-based testing data and processes.

Layer 5: 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 used the four-process delivery system described in Almond, Steinberg, and Mislevy (2002), which is also the conceptual model underlying the IMS/QTI standards in developing a prototype computerized adaptive testing engine that used simulated chemistry laboratory tasks (Seibert, Hamel, Haynie, Mislevy, and Bao, 2006). By parsing assessment systems in terms of the four-process delivery system, the assessment designer is able to describe the computer-based testing procedures, as well as paper- and-pencil tests, informal classroom tests, or tutoring systems. A common 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 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 students or other processes in data structures that are specified in the CAF (e.g., work products, values of observable variables).

Assessment delivery is represented as four principal processes: activity selection, activity presentation, evidence identification, and evidence accumulation. 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 *activity 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 the Assessment Implementation layer of ECD, many assessment delivery systems exist (e.g., Vendlinski, Niemi, Wang, & Monempour, 2006; Wills, et al., 2009 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 Minnesota Comprehensive Assessment-II, Science Assessment

The NSF-funded project, "Application of Evidence-Centered-Design to a State's Large-Scale Science Assessment" (Haertel & Mislevy, 2008a) is supported within the DR K-12 initiative. The project is designed to explore opportunities to leverage principles and structures from ECD in the context of the *Minnesota Comprehensive Assessment, Science Assessment (MCA-II)*, in the area of middle-school science.

A high level overview of the *MCA-II* assessment development process is depicted in Table 3. It begins with Storyboard Development and culminates in the Operational Test Administration. This development process is informed by Minnesota Department of Education's (MDE's) Guidelines for Test Construction and *MCA-II* Test Specifications for Science.

Storyboards are authored by Minnesota science teachers under the direction of Pearson science content specialists. Newly written storyboards are then subjected to Pearson's internal development process, which includes art creation, content verification, editing and formatting, and other internal reviews. Following Pearson's internal process, storyboards are reviewed by MDE content and bias advisory panels. Once reviewed and revised according to MDE recommendations, storyboards are selected for development (Table 3, Stage 1).

Like storyboards, items are authored by Minnesota science teachers for the selected storyboards and are processed by Pearson prior to a second round of reviews by MDE content and bias advisory panels. Items are modified, if necessary, and then may be selected for field-testing (Table 3, Stage 2).

The selection of field test items initiates the electronic development process. Selected storyboards and items are converted to electronic format and prepared for delivery in the testing engine. This task includes item programming and animation development. Another MDE advisory panel reviews the electronic items and necessary adjustments are made (Table 3, Stage 3).

Field test scenarios/items are embedded into the operational test for administration. Following test administration, extended constructed response (ECR) and short constructed response (SCR) items are reviewed by a range finding advisory panel. All tested items are then scored, psychometrically analyzed, and reviewed by a data review advisory panel. (Table 3, Stage 4).

Items are then selected for inclusion on the operational test, after which items are released for public review or reserved for future operational tests (Table 3, Stage 5).

Table 3: Overview of the MCA-II: Science Assessment Development Process

Development Stage	Assessment Development Activity
1	**Storyboard Development** - Authored by MN science teachers and processed by Pearson - Reviewed by MDE content and bias advisory panels - Selected for development
2	**Item Development** - Authored by MN science teacher and processed by Pearson - Reviewed by MDE content and advisory panels - Selected for field test
3	**Electronic Development** - Animations/audio created; items converted to electronic format - FR items programmed; all items reviewed in preview tools
4	**Field Test Administration** - Scenarios/items embedded in operational test - ECR/SCR items reviewed by range finding advisory panel - Items scored; reviewed by data review advisory panel - Items selected for operational test
5	**Operational Test Administration** - Scenarios/items embedded in operational test - Items scored and forms equated - Standards set by MDE advisory panel - Items released to public or reserved for future operational use

Leveraging Evidence-Centered Design in the Development of the Minnesota Comprehensive Assessment-II, Science Assessment

During the *MCA-II* project, we continued to refine our understanding of the leverage points that exist in the large-scale assessment development processes (Mislevy, Steinberg, Almond, Haertel, & Penuel, 2003). In this DR K-12 project, leverage points can be defined as those opportunities and processes that can be refined, using the lens and data structures of ECD, in order to streamline Pearson's assessment design, development, and delivery processes.

Table 4 is a representation of the four leverage points that have been identified to date in the

Pearson test development processes as they are related to different components (or "levers") of ECD.

Table 4: Leverage Points in MCA-II Science Assessment Development Process

Development Stage	Leverage Point (Assessment Development Activity)	ECD Lever(s)	ECD Layer(s)
1	Storyboard Development	1) Standards Alignment, 2) Narrative Structures, 2) Design Patterns	1) Domain Analysis, 2) Domain Modeling
2	Item Development	1) Design Patterns, 2) Task Templates	1) Domain Modeling, 2) Conceptual Assessment Framework
3	Electronic Development of Assessment Items and Forms	1) Task Templates	1) Conceptual Assessment Framework

Standards Alignment

In year 1 of the project the alignment between the NSES Science Inquiry Standards and Unifying Concepts and Processes with the Minnesota State Science Standards was examined. This alignment exercise provided information on the overlap between the two sets of standards, as well as differences in coverage. This information was used to identify a set of topics for design patterns that were used to guide the development of the Minnesota storyboard and item writing. As such, the standards alignment activity was considered a preliminary and necessary step in order to be able to successfully leverage ECD design patterns in the *MCA-II* design and development process.

The decision to align the Minnesota state standards with the Unifying Themes was based on the need to assist assessment designers in conceptualizing ways to assess topics for which it is difficult to create storyboards and items. These topics tend to be more conceptual and higher-level aspects of science—exactly those aspects of science reasoning that were highlighted in the NSES documents.

The agreement between the two sets of standards was based on the rating of features of each set of standards. The rating categories included: "goodness" of alignment between specific Minnesota benchmarks and the focal NSES Inquiry and Unifying Process Standards, and the

extent to which the benchmarks and standards were prioritized in the *MCA* test specifications. A system was developed to rate the goodness of alignments between Minnesota middle school science benchmarks and standards about unifying concepts and processes and middle school-level inquiry from the National Science Education Standards. The criterion employed was clarity of alignment at the detail level. A rating of 3 was met if there was literal matching between key passages; a rating of 2 was met if there was an inferable match at the detail level, and a rating of 1 was assigned if there was an alignment at the broader, higher level only. Content limits from the Minnesota benchmarks were also examined to see if they clarified the alignments, which in some cases they did.

There are several differences between the way the Minnesota benchmarks and National Science Standards are structured that impacted how the alignment review was conducted:

- In the NSES, the middle grades cover 5-8 whereas in the Minnesota benchmarks, they cover grades 6-8. The grade 5 benchmarks are in the grades 3-5 span. Faced with this discrepancy, and in the interest of enabling the broadest examination of alignments across the sets of standards, we looked at Minnesota benchmarks from grades 5 to 8.
- Though the Minnesota benchmarks are grouped into grade span-specific sub-strands, the individual benchmarks are delineated by individual grade. This is in contrast to the National Science Standards, which delineate only by grade spans (K-5, 6-8, 9-12).
- The National Science and Minnesota standard have different hierarchical structures. In the National Standards, the details of each standard are presented in narratives that accompany the standards. In the Minnesota standards, the details appear in benchmarks and content limits. The benchmarks are always relevant for conducting the alignment, yet only those content limits that provide detail about the intent of the benchmarks are relevant. The content limits that are less relevant for alignment review are those that provide possible topics or formats for test items.

To date, benchmarks from the different sub-strands of the MASS History and Nature of Science strand have been reviewed for alignment. Of these, only benchmarks that were identified in the test specifications as appropriate to assess in the *MCA-II* context were reviewed for alignment. Of the 22 benchmarks that fit into this testable category, eight were in the Scientific World View sub- strand, nine in the Scientific Inquiry sub-strand, and eight in the Scientific Enterprise sub-strand. Of the 28 alignments identified among main ideas expressed in the MASS and the NSES, 11 were to the NSES Unifying Concepts and Processes and 17 to the NSES Inquiry standards (Haertel & Mislevy, 2008b).

Narrative Structures

Narrative structures (Fulkerson, Nichols, Haynie, & Mislevy, 2009) are recurring structures for organizing the flow of information and items in the contextualized sets of items that constitute an *MCA-II* task. The task of identifying Narrative Structures was one activity motivated by this analysis of leverage points.

The *MCA-II* test is scenario-based, and the development process begins with the writing of

storyboards. Storyboards are precursors to scenarios and items, serving as the context to which standards-aligned items will be associated. Storyboards describe series of events or natural phenomena, thereby creating real-world contexts for assessment tasks. They are organized into four or five scenes, with each scene consisting of script text and art description that supports the assessment of one or more *MCA-II* Science benchmarks.

Largely unbeknownst to the storyboard writers, each newly written storyboard was implicitly based on one of six primary Narrative Structures. Narrative Structures were initially identified as undergirding components of storyboards during the development of design patterns for the *MCA-II*. During a review of existing *MCA-II* storyboards, a number of storyboards were examined for literary structure and flow and categorized by common narrative features. Each category was then assigned a descriptive name and a characteristic definition. Six categories were identified, with each category representing a unique Narrative Structure (see Table 5).

Table 5. Six Narrative Structure Categories

	Definition
General to specific or whole to parts	A general topic is initially presented followed by the presentation of specific aspects of the general topic.
Specific to general or parts to whole	Specific characteristics of a system or phenomenon are presented, culminating in a description of the system or phenomenon as a whole.
Investigation	A student or scientist completes an investigation in which one or more variables may be manipulated and data is collected.
Topic with examples	A given topic is presented using various examples to highlight the topic.
Change over time	A sequence of events is presented to highlight sequential or cyclical change in a system.
Cause and effect	An event, phenomenon, or system is altered by internal or external factors.

Narrative structures were developed for use by the teams of science teachers who draft storyboards. Once recognized and described, these Narrative Structures could be distributed to storyboard writers for use during the storyboard writing process, thereby potentially increasing the efficiency of storyboard development by improving initial storyboard quality and/or reducing storyboard creation time.

Narrative structures encapsulate experience from previous cycles of *MCA-II* item writing and insights from professional test developers, to help the teacher teams get started

conceptualizing the kinds of storylines needed to develop the thematically-related computer-interactive tasks that comprise the *MCA-II*. Consistent with ECD principles, narrative structures make explicit, and provide a common language for, ways of thinking that the best practitioners have developed less formally and less explicitly. Narrative structures are applied at the Domain Modeling layer of ECD and promote the ECD principles of generativity, re-use, and explicitness with regard to task conceptualization. Narrative structures are a new representational form created in this project; similar to the work of Georges Polti (1916), The thirty-six dramatic situations. Polti presented 36 plots that serve as patterns for writers developing storylines. Likewise, the six narrative structures identified in support of the creation of storyboards, serve as patterns for storyboard writers to use as they develop the storyline that moves through the *MCA-II* scenarios.

Prior to recognizing that Narrative Structures are inherent components of the storyboard development process, storyboard writers were asked to rely on their own methods and tools to organize thoughts, develop themes, and construct outlines. These tools may include trial-and-error, concept mapping, outlining, graphic representations, and other methods of organization. Essentially, storyboard writers are assigned a task and asked to create a storyboard without any significant direction. This condition often resulted in ambiguity, frustration, and inefficiency on the part of the storyboard writers. By recognizing, explicating, and distributing Narrative Structures to storyboard writers for use as advance organizers, efficiencies can be gained and frustration can be reduced.

In order to ascertain the usefulness and impact of narrative structures on the storyboard and item-writing process, a study was conducted in January 2008 where six narrative structures were presented to half of a group of 16 writers attending a storyboard writing training workshop. These eight writers were instructed to use the narrative structures as tools to facilitate their independent storyboard writing process - applying the narrative structures as an organizer during the brainstorming, writing, and revising stages of storyboard development. Writers from both the narrative structure and control groups were then asked to provide feedback via information sheets, an on-line survey, and focus group interviews (Haertel & Mislevy, 2008b).

Design Patterns

Design patterns (Mislevy et al., 2003) are knowledge structures created in the PADI Project, which are now being tuned for use in the DR K-12 project (see Figure 1 for screenshot design pattern template in PADI system). During Year 2, the design pattern form was tailored to the needs of the *MCA-II*. Structured around the form of assessment arguments and focusing on particular aspects of knowledge or skill, design patterns provide scaffolding that task designers can use to build scenarios and items to assess those aspects of knowledge and skill. The original PADI Project focused on creating design patterns to guide the design of tasks for science inquiry.

We continue this line of work in the current project by developing design patterns for hard to assess science topics (e.g., conducting observational or experimental investigations; see Table 6), as well as a suite of design patterns organized around the concept of model-based reasoning (see

Table 7). We have identified storyboarding and item design and development to be leverage points in the Pearson assessment development processes that we can impact with design patterns. The design patterns developed in the PADI Project served as suitable starting points for the present design pattern work. We anticipate that the newly developed design patterns will support the *MCA-II* and improve the efficiency of storyboard and item development, assist writing teams address hard-to-assess benchmarks, and make explicit the validity argument. Being able to view individual storyboards and items as instances motivated by design patterns should help the Pearson review teams in their analyses of items before they move to production and after they are implemented electronically.

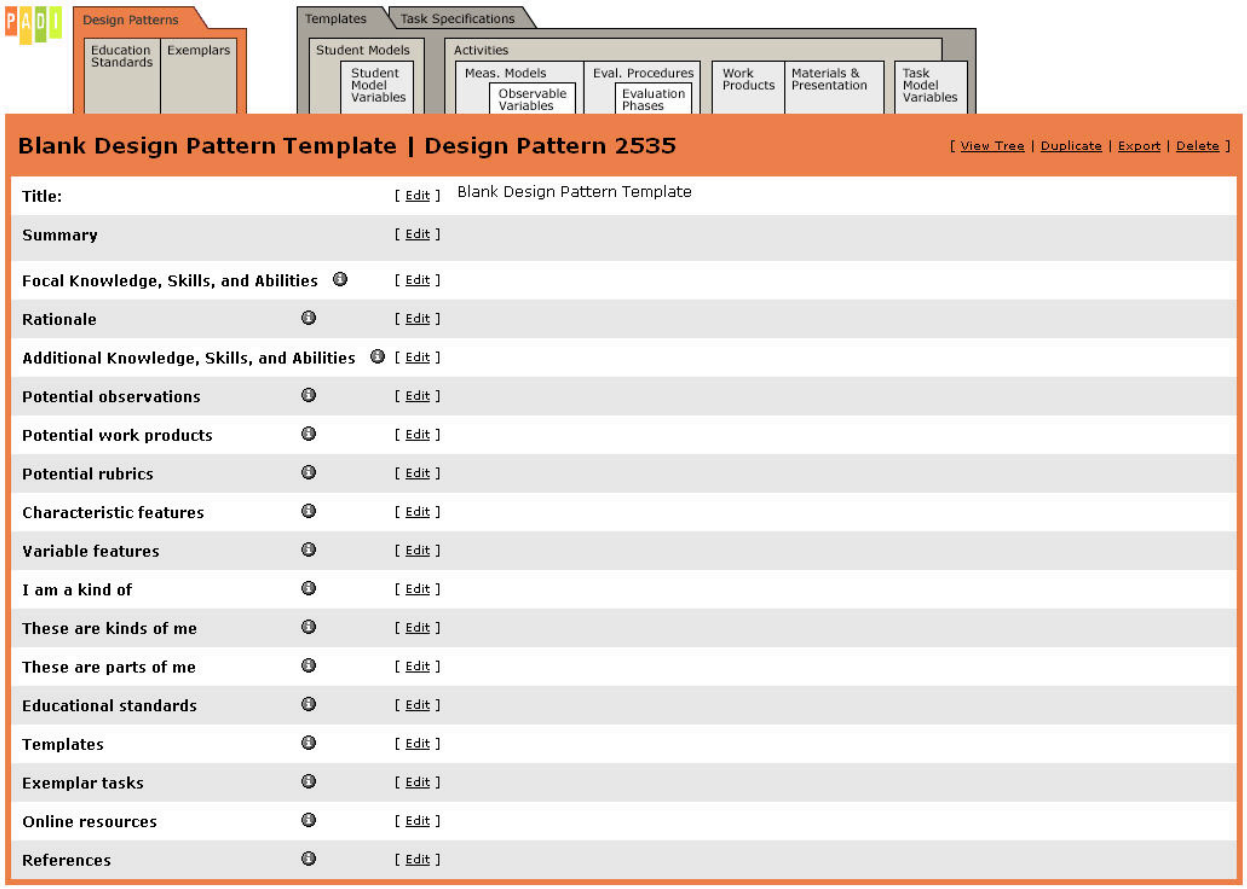

Figure 1. Design Pattern Template

Table 6. Design Patterns for Hard to Assess Science Topics

Title	Overview
Observational Investigation	This design pattern supports the writing of storyboards and items that address scientific reasoning and process skills in observational (non-experimental) investigations. Observational investigations differ

	from experimental investigations. In experimental investigations, it is necessary to control or manipulate one of more of the variables of interest to test a prediction or hypothesis; in observational investigations, variables typically cannot be altered at all (e.g., objects in space) or in a short time frame (e.g., a lake ecosystem). This design pattern may be used to generate groups of tasks for any science content strand.
Experimental Investigation	This design pattern supports the writing of storyboards and items that address scientific reasoning and process skills in experimental investigations. In experimental investigations, it is necessary to manipulate one or more of the variables of interest and to control others while testing a prediction or hypothesis. This contrasts with observational investigations, where variables typically cannot be manipulated. This design pattern may be used to generate groups of tasks for science content strands amenable to experimentation.
Reasoning about Systems and Complexity	This design pattern supports the writing of storyboards and items that address reasoning within the context of complex systems. Complex systems are characterized interactions among components of the system and, typically, outcomes that emerge without any explicit driving force. Tasks targeting these types of systems typically require multi-step causal reasoning and the consideration of the effects of multiple concepts or factors within a system. The prevalence of complex systems across scientific content domains suggests the development of a design pattern that enables design of tasks that target students' reasoning about complex systems, across domains and, as a result, grade levels.
Genetics Learning Progression for Grades 5-10	This design pattern describes students' evolving knowledge of the characteristics and functions of genes. Its contents are based on a published journal article by Duncan, Rogat, and Yarden (2009) in which the authors posit a learning progression for deepening students' understandings of modern genetics across grades 5-10. This understanding of modern genetics is identified in the paper as consisting of understanding of the genetic model, the molecular model, and the meiotic model. The paper posits eight main ideas about the three models followed by the characteristics of what students in grade bands 5-6, 7-8, and 9-10 are capable of understanding about the main ideas respectively.
Model Use in Interdependence Among Living Systems	This design pattern supports developing tasks that require students to reason through the structures, relationships, and processes of ecological models. Use of ecological models is often combined with the formation of ecological models in tasks. Many tasks that address evaluation and revision of ecological models also involve the use of these models.

Table 7. Suite of Design Patterns for Model-Based Reasoning

Title	Overview
Model Articulation	Tasks supported by this design pattern assess student's ability to articulate the meaning of physical or abstract systems across multiple representations. Representations may take qualitative or quantitative forms. This DP is relevant in models with quantitative and symbolic components (e.g., connections between conceptual and mathematical aspects of physics models) Model articulation is often be pertinent in multiple-step tasks, after the model formation step.
Model Elaboration	This design pattern supports developing tasks in which students elaborate given scientific models by combining, extending, adding detail to a model, and/or establishing correspondences across overlapping models. This design pattern can considered a special case of model formation in that the aim is to develop a modeled conception of a situation. But the emphasis is what is happening in the model layer with respect to extensions of models or connections between models. Model elaboration is also similar to model revision, in that a given model or a set of unconnected models does not account properly for the target situation and reformulation is required.
Model Evaluation	This design pattern supports developing tasks in which students evaluate the correspondence between a model and its real-world counterparts, with emphasis on anomalies and important features not accounted for in the model. This design pattern is tied closely with model use, and is also associated with model revision and model elaboration.
Model Formation	This design pattern supports developing tasks in which students create a model of some real-world phenomenon or abstracted structure, in terms of entities, structures, relationships, processes, and behaviors. The Model Formation design pattern can be viewed as a subpart of the Model-Based Inquiry design pattern, and many tasks combine Model Formation with Model Use. The Model Formation design pattern also overlaps with those for Model Elaboration and Model Revision.
Model Revision	This design pattern supports developing tasks in which students revise a model in situations where a given model does not adequately fit the situation or is not sufficient to solve the problems at hand. Because its centrality, model revision is difficult to assess in isolation from other aspects of model-based reasoning. Model revision is prompted only by model evaluation, and then model formation must be used to propose alternatives or modifications.
Model Use	This design pattern supports developing tasks that require students to

	reason through the structures, relationships, and processes of a given model. Model use is often combined with model formation in the same tasks, and most tasks that address model evaluation and model revision also involve model use.
Model-Based Inquiry	This design pattern supports developing tasks in which students work interactively between physical realities and models, using principles, knowledge and strategies that span all aspects and variations of model-based reasoning.

In 2009, a second pilot study was conducted by Pearson to gather some initial information about the use of design patterns on storyboard and item development tasks. Conducted in April, this study involved 10 writers who had an average of 4.2 years of experience at Pearson and an average of 18.5 years teaching. All 10 writers participated in a brief workshop and then were given one hour for a writing assignment in which they were given 3-5 benchmarks and asked to create a storyboard of at least 4 scenes and to write one related multiple-choice item. Six of the writers were given a short training session on the task itself plus 30 minutes of training on design patterns. Each member of this group (the DP condition) received either the Observational Investigation Design Pattern or the Experimental Investigation Design Pattern and was told that he or she had the option of using the design pattern for the assignment. The remaining four writers (the non-DP condition) received training on the assigned task but no training about or access to design patterns. All writers were asked to communicate their experiences with the assignment via an on-line survey and a one-hour focus group. Three methodologies were used in the pilot study: protocol analysis, focus groups, and a survey (Haertel & Mislevy, 2009).

It is important to note that the design patterns that were developed for this project are cast at a level that is useful for curriculum developers and classroom teachers in Minnesota and beyond. Recall that the design patterns are based on the Minnesota Academic Standards for Science (MASS), which are strongly connected to NSES standards and unifying themes, as well as Project 2061. Thus, the design patterns could potentially assist the curriculum developers and teachers to think through the design of quality assessments in these hard-to-assess, but critical areas of science (e.g., model-based reasoning, systems thinking) and, in doing so, better integrate large-scale testing practices with instructional practice.

Task Templates

Task templates, like design patterns, were also developed as part of the PADI assessment design system (Riconscente, et al., 2005). Task templates are a supporting representation for more technical aspects of task design and implementation, and address issues such as form and expression of scoring algorithms, specifications for presentation material and work products, and presentation logic. These templates support the expression of these elements in IMS/QTI standards for electronic assessment systems. They promote re-use in implementation and delivery.

It is worth noting that the use of task templates leverages ECD ideas in a distinct way from design patterns. Design patterns focus on conceptual and substantive aspects of assessment design. They help a designer build tasks around key ideas in some substantive area, such as model-based reasoning or experimental investigations, and decide what features to include in tasks, what kind of work products to capture, how to score student work, and so on. Validity and efficiency of the intellectual aspects of the design process are at issue here, regardless of the form of the assessment, the mechanics of implementation, or the details of delivery.

The conceptual aspects of design decisions are already built into task templates. Task templates focus on efficiencies in implementing tasks, by using elements of scoring code, modules of psychometric models, presentation material and interface elements, and so on, in order to maximize interoperability and re-use. They provide mechanical rather than conceptual leverage.

As the project proceeds, we will analyze the data structures underlying the Pearson/MDE task types for opportunities to streamline the work of the Pearson task implementers and assist in shaping the work of the task designers around task and evidence structures. Task designers may benefit from the use of task templates as they provide conceptual frames and support interfaces to organize their work in ways that segue easily into implementation in existing schemas for presentation and response evaluation. The routines for implementing, rendering, and scoring figural response tasks, for example, provides opportunities for formalization and added efficiency through the use of templates.

We will begin our work on Task Templates in Year 3 of the project. By their nature, task templates are more specific than design patterns to delivery systems and operational processes. Scaling up templates is beyond the scope of the present work, so the contributions in the current project on task templates will be more limited in scope. We anticipate developing one or two templates and supporting discussion to illustrate the ideas and provide a small contribution to operational work.

Summary & Discussion

The purpose of this paper is to highlight how evidence-centered assessment design (ECD) can be leveraged to improve validity and maximize efficiency in large-scale test design, development and delivery. We began by describing the five layers of ECD (see Table 1) that test designers can use for explicating the measurement domain and relevant constructs (Domain Analysis and Modeling); structuring assessment elements such as tasks, rubrics, and psychometric models into an assessment argument (Conceptual Assessment Framework); and for implementing these elements in an operational assessment (Assessment Implementation and Delivery). Next, we introduced current work applying ECD to the design and development of the *Minnesota Comprehensive Assessment-II, Science Assessment (MCA-II)*, and provided a high level overview of the primary stages of the *MCA-II* development process. Finally, we highlighted particular features of ECD – standards alignment, narrative structures, design patterns and task templates – that we have identified as ways to improve validity and efficiencies in the *MCA-II* assessment design, development, and delivery processes.

As we continue with the *MCA-II* project we will further refine our understanding of the roles of standards alignment, narrative structures, design patterns and task templates, as well as identify novel ways in which ECD can be leveraged, in helping improve the validity and efficiency of the large-scale assessment development process. It was clear from the beginning of the project that a detailed understanding of the development and delivery processes was essential to leveraging ECD ideas. As described in Section 4, this analysis was carried out, and provided a foundation for the work done thus far, as well as work currently under way. However, a key insight in our work to date has been the important of understanding the psychology of storyboard and item-writing processes. Simply providing additional information and structure to test developers in the form of narrative structures and design patterns is not sufficient to make it useful. Rather, understanding the complex, iterative, creative, and constrained design challenge of storyboard development in particular proved necessary to making the storyboards useful. A project technical report by Haertel et al. (forthcoming) describes the ways that the interviews with users of design patterns spurred the development of representation forms and interactive tools for developers to use design patterns, and the approach for practice and training in how to integrate them with their other design aides.

We believe that evidence-centered assessment design also has potential to be leveraged to improve the validity and efficiency of the development of classroom assessment systems. We point in particular to the design patterns we have developed here. As discussed in Section 4.2, the design patterns we have created to support storyboard and item writing for the *MCA-II* address important, yet hard-to-assess, aspects of science, including model-based reasoning, systems thinking, and experimental and observational investigation. Although we chose these areas to be particularly helpful in developing the *MCA-II*, we cast them at a level that would also support the development of classroom and curriculum-based assessments.

The idea here is that task development would center around the science and associated concepts rather than the forms of assessment and the constraints under which a particular assessment has to operate—interactive and low-stakes in the classroom, for example, but constrained and concise in the *MCA-II*. The design patterns, by virtue of their organization around science concepts rather than item types, promote a better integration between the large-scale assessment and learning assessment. Designers, whether they are *MCA-II* item writers or classroom teachers, can see the tasks they create as addressing the same underlying science, in terms of common Focal KSAs, Characteristics Features of Tasks, and Potential Observable features of students' performances. Different choices about Additional KSAs, Variable Features of Tasks, and Work Products are required in order to meet the varying constraints and purposes of different assessment contexts. Having common and explicit design patterns thus enhances the instructional validity of assessment as well as the evidentiary value of tasks.

Although not an explicit focus of the current project, we believe that the student and evidence models (part of the CAF layer of ECD) can also be leveraged to improve the validity and efficiency of classroom assessment systems. Recall that 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 finer-grained student and evidence models.

For example, a student model for the Verbal Reasoning domain on the GRE has been represented as a single unobservable variable representing proficiency in that domain and a probability distribution across the range of values the variable might take. However, representing verbal reasoning ability as one variable (e.g., understanding word meaning) does not represent the target construct in a manner that provides teachers with the details they need to give students formative feedback regarding several aspects of their verbal reasoning skills. Instead, a student model for verbal reasoning could consist of several variables (e.g., understanding word meaning, evaluating word use in context, drawing inferences and conclusions from text, etc.), each representing a different aspect, or even developmental stage, of proficiency in that domain. Specifying a student model at this level gives teachers the ability to relate assessment performance back to several aspects of verbal reasoning proficiency and diagnose learning challenges at a finer grained level of detail.

Likewise, an evidence model for the Verbal Reasoning domain on the GRE has been represented as consisting of whether each item is answered correctly or incorrectly (i.e., the observable variable) and a unidimensional measurement model that links the student model with the observable variable. As with the larger-grained student model, however, limiting evidence of verbal reasoning ability to correct or incorrect responses does not give teachers the type of evidence they need to provide students with diagnostic feedback. Instead, an evidence model for verbal reasoning could consist of several observable variables (e.g., response accuracy, complexity, relevance, efficiency, etc.), each representing a different type of evidence of verbal reasoning proficiency. Specifying an evidence model at this level gives teachers the ability to use multiple sources of evidence to pinpoint areas where students are especially strong or deficient in their verbal reasoning skills.

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, *1*(5). Retrieved May 20, 2007 from http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

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 (CRESST), 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.

Duncan, R.G., Rogat, A.D., Yarden, A. (2009) A Learning Progression for Deepening Students' Understandings of Modern Genetics Across the 5th–10th Grades. *Journal of Research in Science Teaching*, 46(6), 655-674.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

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

Fulkerson, D., Nichols, P., Haynie, K., & Mislevy, R. (2009). *Narrative structures in the development of scenario-based science assessments* (Large-scale Assessment Technical Report 3). Menlo Park, CA: SRI International.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns.* Reading, MA: Addison-Wesley.

Gorin, J. (2006). Test design with cognition in mind. *Educational measurement: Issues and practice, 4*(25), pp. 21-35.

Haertel, G. D., & Mislevy, R. J. (2009). Year 2 annual report for Application of Evidence-centered Design to a State, Large-Scale Science Assessment. (NSF DRL-0733172). Menlo Park, CA: SRI International.

Haertel, G. D., & Mislevy, R. J. (2008a). Year 1 annual report for Application of Evidence-centered Design to a State, Large-Scale Science Assessment. (NSF DRL-0733172). Menlo Park, CA: SRI International.

Haertel, G., & Mislevy, R. (2008b). Application of evidence-centered design in large-scale assessment (NSF DRL-0733172).

Huff, K., & Plake, B. (2009). Evidence-centered design in practice. Coordinated session presented at the annual meeting of the National Council on Measurement in Education (NCME), San Diego, CA.

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 validity. *Psychological Bulletin, 112,* 527-535.

Leighton, J.P., & Gierl, M.J. (Eds.) (2007). *Cognitive diagnostic assessment for education: Theory and practices*. Cambridge University Press.

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 (NCME), New Orleans, LA.

Messick, S. **(**1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2)*,* 13-23.

Mislevy R. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469.

Mislevy, R., & Riconscente, M. M. **(**2006). Evidence-centered assessment design: Layers, structures, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90)*.* Mahwah, NJ: Erlbaum.

Mislevy, R. & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), pp. 6-20. Malden, MA: Blackwell Publishing.

Mislevy, R., Steinberg, L., Almond, R. G., Haertel, G. D., & Penuel, R. (2003). *Leverage points for improving educational assessment (PADI Technical Report 2).* Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives 1*, 3-67.

Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., et al. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report 1). Menlo Park, CA: SRI International. Retrieved May 20, 2007 from http://padi.sri.com/downloads/TR1_Design_Patterns.pdf

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.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*, 485-504.

National Research Council (2001). Advances in the sciences of thinking and learning. In J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.), *Knowing what students know: The science and design of educational assessment* (pp. 59-110). Washington, D.C.: 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). Reconceiving 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.

Polti, Georges. (1916). *The Thirty-Six Dramatic Situations*. Boston: The Writer, Inc.

Quellmalz, E., Timms, M., & Schneider, S. (2009). *Assessment of Student Learning in Science Simulations and Games*. WestEd.

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.

Seibert, 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 (PADI Technical Report 15).* Menlo Park, CA: SRI International.

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.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Vendlinski, T.P., Niemi, D.M., Wang, J., & Monempour, S. (2006). Improving formative assessment practice with educational information technology. *Journal on Systemics, Cybernetics and Informatics, 6,* 27-32.

Vendlinsky, T. P., Niemi, D., & Baker, E. L. (2008). 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.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.) (2006). *Automated scoring of complex tasks in computer based testing*. Mahwah, NJ: Erlbaum Associates.

Wills, G., Davis, H., Gilbert, L., Hare, J., Howard, Y., Jeyes, S., Millard, D. and Sherratt, R. (2009) Delivery of QTIv2 question types. *Assessment & Evaluation in Higher Education*, 34, 353-366.

Wilson, M. R. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.