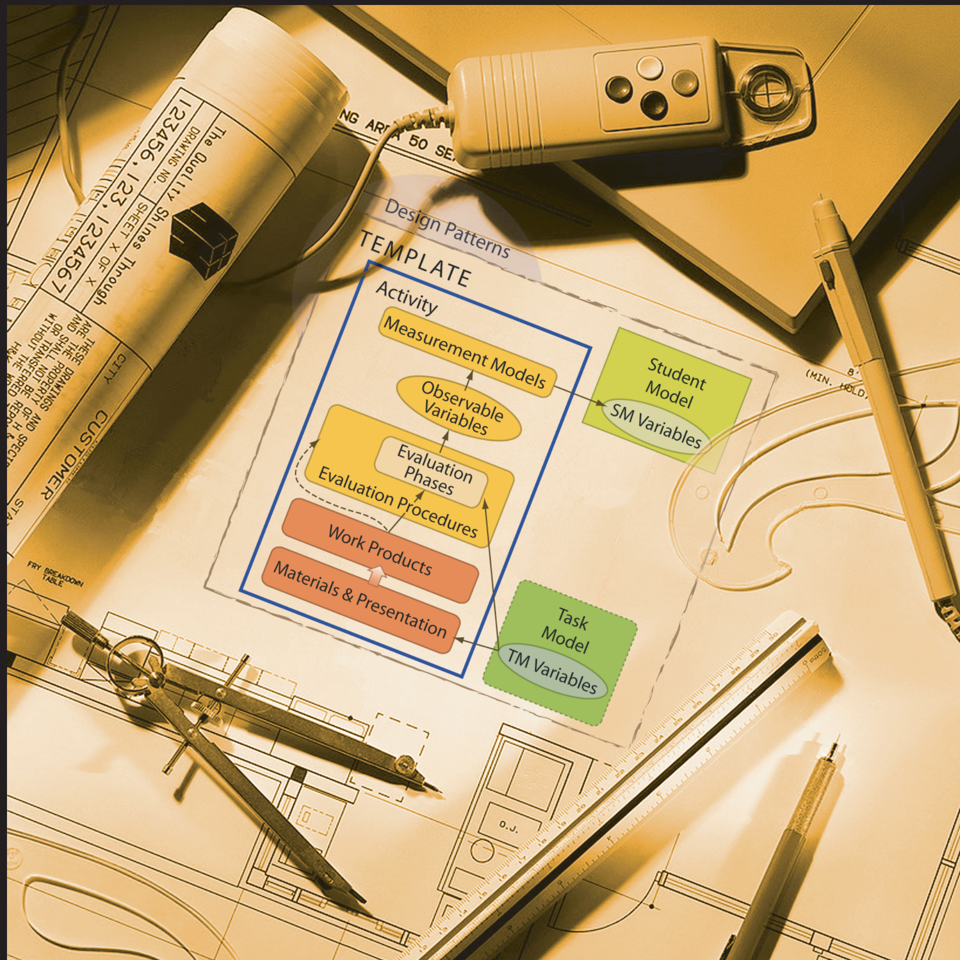

PADI Technical Report 5 | May 2005

The Case for an Integrated Design Framework for Assessing Science Inquiry

PADI | Principled Assessment Designs for Inquiry

Gail P. Baxter, Mathematica Policy Research, Inc.
Robert J. Mislevy, University of Maryland

Report Series Published by SRI International

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

PADI Technical Report Series Editors
Alexis Mitman Colker, Ph.D. *Project Consultant*
Geneva D. Haertel, Ph.D. *Co-Principal Investigator*
Robert Mislevy, Ph.D. *Co-Principal Investigator*
Klaus Krause. *Technical Writer/Editor*
Lynne Peck Theis. *Documentation Designer*

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

PRINCIPLED ASSESSMENT DESIGNS FOR INQUIRY
TECHNICAL REPORT 5

The Case for an Integrated Design Framework for Assessing Science Inquiry

Prepared by:
Gail P. Baxter, Mathematica Policy Research, Inc.
Robert J. Mislevy, University of Maryland

Acknowledgments
PADI is supported by the Interagency Educational Research Initiative (IERI) under grant REC-0129331 (PADI Implementation Grant). Naomi Chudowsky and Alissa Morrison contributed to the construction of the design patterns in the appendix.
Disclaimer
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

CONTENTS

FIGURES

Abstract

In this report, we provide a rationale and approach for articulating a conceptual framework and corresponding development resources to guide the design of science inquiry assessments. Important here is attention to how and why research on cognition and learning, advances in technological capability, and development of sophisticated methods and techniques in measurement can and should be put to use in designing maximally informative assessments. To ensure quality and continuity in the design process, the framework advocates an evidence-centered approach in which the components of assessment design (i.e., substantive arguments, design elements, and operational procedures) are described and their relationships elaborated. Further, assessment-design data structures expressed in terms of extensible object models (i.e., reusable parts) and supported by Web-based tools, facilitate generating, exchanging, and reusing particular components of the design process. A shared, practical, and instructionally informative set of assessment design tools, both conceptual and computer based, can serve to speed the diffusion of improved assessment practices.

1.0 Introduction

The past decade has witnessed considerable activity aimed at bringing assessment practices in line with goals for learning and attendant changes in curriculum and instruction. Progress has been made, for example, in embedding assessments in technology-supported learning environments, creating complex performance-based tasks, tracking student reasoning during problem solving (e.g., strategy use, metacognition), and evaluating multiple aspects of student performance or products over time. However, much of this work has been localized or experimental in nature and generally not cost-effective, not easily adaptable for large-scale use, and not reusable for other purposes or in other contexts. As such, research and development have produced little in the way of a shared, practical, and instructionally informative set of tools and strategies to assess learning. What is needed is an integrated framework that coordinates but does not constrain assessment design, a framework that capitalizes on previous efforts while providing a generalized but principled and coherent approach to guide future efforts. In this report, we present a rationale and approach for explicating such a framework for assessing science inquiry.

The formulation of an integrated assessment design framework is made possible by the coalescence of three lines of research and development (Mislevy, Steinberg, & Almond, 2002; Pellegrino, Chudowsky, & Glaser, 2001). First, current understandings of how students acquire and use knowledge serve to identify appropriate targets of assessment and denote the nature of evidence that should be elicited. Second, improvements in technological capabilities enable the administration of assessment tasks that mirror the complexity of inquiry learning and facilitate the collection and evaluation of data to support standards-based claims about student knowledge/understanding. Third, advances in measurement methods and statistical techniques make it possible to simultaneously weigh multiple aspects of student performance and attend to the influence of contextual factors when establishing the validity of claims or inferences about student knowledge or understanding. Taken together, these developments provide the essential underpinnings for a practical and feasible assessment design framework, one in which the components of assessment design (i.e., substantive arguments, design elements, and operational procedures) are described and their relationships elaborated.

Here we focus on a design framework for assessing science inquiry being developed by The Principled Assessment Designs for Inquiry (PADI) project, an NSF-sponsored collaboration among researchers and developers at SRI International, the University of Maryland, the University of California-Berkeley, the University of Michigan, and the Lawrence Hall of Science. The framework makes explicit the links between educational standards and curricular goals on the one hand and assessment tasks and score criteria on the other. Second, the framework provides guidance for the development of high-quality assessments in the form of *design patterns* and *task templates* expressed in terms of extensible object models.[1] Third, the framework unifies the elements of assessment design, delivery, and evaluation to help a developer ensure that critical considerations (e.g., consistency, usability, validity) inform the process from its inception. In what follows, we describe the multidisciplinary approach taken by PADI to conceptualize an assessment

[1] The reader is referred to Rumbaugh, Jacobson, and Booch (1998) for an overview of an object modeling approach to software design and the application of these ideas to modeling business or other systems.

design framework and a collection of development resources for designing assessments of science inquiry.

We begin with a brief review of three contributing developments that make possible the formulation of a practical, conceptually grounded assessment design framework: research on cognition and learning, advances in technological capability, and the availability of increasingly sophisticated methods and techniques in measurement. The first of these developments, concerning the nature of learning, is foundational. By itself it opens the door to improving assessment, whether or not specific technologies or measurement models are pertinent to a given assessment use.[2] By making underlying theories of learning explicit in the PADI framework, one can translate educational goals effectively into assessment tasks and appropriate score criteria. The second and third developments—technology and measurement—support the valid and reliable assessment of multifaceted inquiry in meaningful contexts. Conventional assessment approaches address content knowledge, specific process skills, and some aspects of science inquiry (e.g., analysis and interpretation of data) fairly well. Less satisfactory are efforts to develop assessments that exemplify the essence of science inquiry—interactive, cyclical, and constructive—this despite the importance given to inquiry in standards documents and curricular materials. In our view, a much closer alignment of assessment with the complexities of inquiry teaching and learning can be realized through the use of innovative technology (to deliver and score assessments) and powerful measurement methods (to summarize and interpret performance).

Next, we describe the key features of the PADI assessment design framework. In particular, we emphasize the centrality of an evidence-centered approach to assessment design, an approach that is guided by four critical questions: (a) What does it mean to know and do inquiry? (b) What constitutes evidence of knowing? (c) How can that evidence be elicited from students? (d) What are appropriate statistical techniques for making valid inferences about what students know from what students do? Second, we describe two data structures—*design patterns* and *task templates*—structures that guide assessment designers through the elements of evidence-centered design. A *design pattern* describes, at a conceptual level, common and unique features of families or sets of science inquiry assessments and bridges the content expertise and measurement expertise needed to create usable and useful assessments. *Task templates* encompass the technical considerations necessary to move from the substantive foundation (expressed in narrative fashion in *design patterns*) to specifications for particular tasks and the operational processes necessary to carry out the assessment. Third, we comment on the use of object modeling, a software design strategy, to develop Web-based structures (i.e., PADI *design patterns* and *task templates*) composed of reusable parts. Formulating PADI structures as design objects facilitates generating, sharing, and reusing elements of the design process and circumvents a "from-the-ground-up" approach to assessment development. To conclude, we comment briefly on the ongoing development of a scoring engine and the creation of exemplar tasks.

[2] Informal classroom observations may not require technology or measurement models, whereas computer-based coached practice systems require both. Large-scale, high-stakes tests may involve technology, sophisticated measurement techniques, or both.

2.0 Contributing Developments

Three messages in the National Research Council's report *Knowing What Students Know: The Science and Design of Assessment* (Pellegrino et al., 2001) serve to situate the PADI effort. First, current conceptions of student cognition and how people learn combined with goals for science learning (e.g., American Association for the Advancement of Science [AAAS], 1993; National Research Council, 1996) provide the substantive underpinnings for the design and interpretation of assessments. Second, technology enables the administration of complex and realistic tasks and the accumulation of direct evidence of student thinking, reasoning, or understanding. Third, measurement or statistical models make possible the integration and interpretation of multiple pieces of information to support valid inferences about what students know and can do. In what follows, we elaborate on each of these messages and the consequent opportunities for, and challenges to, the improvement of assessment design.

2.1 Learning and Cognition

The essential conceptual component for designing educational assessments is the characterization of competence within a subject matter. Psychological research on learning and cognition has, at various points in time, emphasized different aspects of knowing, understanding, and reasoning. In the last 40 years, the cognitive perspective (with its emphasis on knowledge structures) and the situative perspective (with its emphasis on social situations) have presented views of achievement that have challenged the principles underlying extant teaching practice and test design. The history of developments in these and other areas is described by Greeno, Pearson, and Schoenfeld (1996). Here we present a brief description of the cognitive and situative perspectives.

The cognitive perspective focuses on structures and uses of knowledge, including principles and concepts of subject matter domains, the organization of information (schemas, mental models), and procedures and strategies for problem solving and reasoning (e.g., Anderson, 2000). Studies of expertise in various domains have demonstrated that the nature and quality of cognitive activity underlying an individual's performance reflects the experience, degree of learning, and state of knowledge of the problem solver (Chi, Glaser, & Farr, 1988; Ericsson & Smith, 1991). The recurring theme is that learning is a process of constructing new knowledge on the basis of current knowledge. As learning occurs, increasingly well-structured and qualitatively different organizations of knowledge develop. Most important is the integration of declarative or factual knowledge with an understanding of when and how to use that knowledge. It is this integrated or connected knowledge that enables certain cognitive activities, such as building a mental model or representation of a problem to guide its solution, managing one's thinking while performing a task, enlisting appropriate goal-directed solution strategies to facilitate problem solving, and generating and elaborating explanations. Because observable differences in these cognitive activities—problem representation, metacognition, strategy use, explanation—are associated with differential levels of understanding, they are appropriate criteria for evaluating student performance and achievement (Baxter & Glaser, 1998).

Whereas the cognitive perspective emphasizes the individual development of knowledge, the situative perspective draws attention to the social and participatory aspects of learning (e.g., Brown, Collins, & Duguid, 1989). From the situative perspective, learning science involves extended experience with, and membership in, a community of people who practice science. To this end, classrooms are structured as communities of collaborative, reflective practice in which students are challenged to think deeply about, and to engage actively in, doing science (e.g., Bruer, 1993). Teachers in these classrooms assume the role of representatives of the scientific community. In this role, they "are expected to model reflection, fostering a learning environment where students review each others' work, offer suggestions, and challenge mistakes in investigative processes, faulty reasoning, or poorly supported conclusions" (National Research Council, 1996, p. 88). These "situated" participatory experiences lead students to pick up certain practices and forms of discourse, adopt certain ways of perceiving the discipline, and encourage habits of mind and particular ways to view the world (Greeno, Collins, & Resnick, 1996).

Important to both the cognitive and situative perspectives is an emphasis on learning with understanding in meaningful contexts. In science education, standards documents and curricular materials promote inquiry as a key strategy for engaging students in learning science. "Inquiry is a multifaceted activity that involves making observations; posing questions; examining books and other sources of information to see what is already known; planning investigations; reviewing what is already known in light of experimental evidence; using tools to gather, analyze, and interpret data; proposing answers, explanations, and predictions; and communicating the results. Inquiry requires identification of assumptions, use of critical and logical thinking, and consideration of alternative explanations" (National Research Council, 1996, p. 23). Engaging in inquiry allows students to experience the ways in which scientists study the world and encourages an understanding of the nature of science and scientific knowledge. Key here is a view of science as an ongoing cyclical process of constructing and modifying ideas, theories, and/or models through the systematic gathering of evidence, application of logical argument, and questioning of assumptions, procedures, and conclusions. As students' experiences with inquiry accumulates, discipline-specific variations in modes of inquiry and canons of evidence give way to unifying concepts and processes that transcend grade and disciplinary boundaries.

Taken together, theories of learning, education standards, and instructional expectations provide the substantive underpinnings for science assessments. They serve to identify (at a general level) relevant goals of assessment and the nature of evidence that should be elicited to support claims or inferences about student understanding or achievement; they are not specifically geared toward guiding assessment design. Well-established procedures for designing traditional assessments, procedures that have evolved over time to ensure consistency and coherence, have proved unsatisfactory, in and of themselves, for designing more complex assessment tasks. Indeed, analyses of "innovative" assessments have pointed to inconsistencies among assessment goals, developed tasks, and/or score criteria (e.g., Achieve Inc., 2002; Baxter & Glaser, 1998; Means & Haertel, 2002). Further, a task-centered approach, characteristic of many efforts to design complex assessments (particularly performance assessments), has resulted in some innovative assessment situations but not necessarily effective strategies for summarizing and drawing inferences from the multiple pieces of information elicited from students. We argue that one must design assessments from the very start around the inferences one wants to make, the observations one needs to group them, the situations that will evoke these observations, and the chain of reasoning that connects them. The central issues are construct definition, forms of evidence, and situations that provide evidence regardless of the means by which data are to be gathered and evaluated (Messick, 1994).

PADI introduces *design patterns* as a tool for structuring substantive considerations into an assessment argument. An assessment argument lays out the chain of reasoning from evidence (what students say or do in particular situations) to inference (what we wish to say about students' abilities more generally). The key elements of an assessment argument—what is important to know, what constitutes evidence of knowing, and in what ways this evidence can be elicited from students—are explicated in *design patterns* (see below for examples). Making substantive considerations explicit from the onset serves to place appropriate boundaries on subsequent design decisions. Because assessment design is inevitably iterative, a process of inquiry itself, design decisions always can be revisited in light of reflection and empirical feedback. The point is to ensure that the designed assessment is (a) consistent with the developer's goals/intentions and (b) internally coherent (i.e., evidence is gathered and interpreted in ways that bear on the underlying knowledge and purposes the assessment is intended to address).

2.2 Technological Developments

Increases in the availability and capability of technology have the potential to positively influence and assist developers and users of assessments. Unlike the paper-and-pencil modalities of conventional large-scale assessments, technology can provide realistic work environments, track students' strategies and progress as they solve problems, and yield rich evidence about students' reasoning processes. In essence, technology permits the grounding of assessment in cognitive conceptions of knowing and facilitates the acquisition of evidence of student understanding more efficiently and effectively than do traditional assessments. In other words, technology provides an infrastructure that enables the delivery and scoring of complex assessments.

In recent years, technology has figured prominently in efforts to design intelligent tutoring systems (e.g., Koedinger & Anderson, 1993), to promote student acquisition of coherent mental models of important subject matter concepts (e.g., Hunt & Minstrell, 1994), to provide frequent opportunities for formative assessment with rich feedback to students and teachers (Barron et al., 1995; 1998; Cognition and Technology Group at Vanderbilt, 1994, 1997), and to emphasize and promote self-assessment and group problem solving (e.g., White & Frederiksen, 1998; 2000). This work is based on cognitive conceptions of what it means to know and learn, and is often combined with sophisticated statistical or psychometric techniques to model the complex performances observed in these situations. Two examples of technology-based assessments—the first developed from a cognitive perspective and the second from a situative perspective—illustrate some of the key ideas.

Drawing on a cognitive perspective, Ron Stevens and his colleagues have developed Interactive Multimedia Exercises (IMMEX) (University of California, Los Angeles, n.d.), an online problem-solving environment predicated on a model of scientific inquiry (e.g., Stevens, Lopo, & Wang, 1996). Each case begins with a descriptive scenario for which students are expected to frame the problem, judge what information is relevant for solving the problem, plan a strategy for searching available information, gather "data," and then draw relevant conclusions. For example, students in environmental science may be asked to determine why dead fish are washing up on the shores of a river. In biology, students may take on the role of forensic scientists in an effort to identify the parents of a girl who suspects she was the victim of a mix-up in the maternity ward. The problem-solving environment is structured to allow students to select from a number of choices (via pull-down menus) what tests to do and the sequence in which to conduct the tests. The software records a student's every step as she/he attempts to solve each case. Patterns in student problem-solving performance are identified, and similar performances are clustered by using the statistical machinery of artificial neural networks (e.g., Vendlinski & Stevens, 2002). From this information, graphs are constructed to display performance change (in terms of strategy use) over time for an individual student and for groups of students. Consistent with the literature examining differences between experts and novices, Stevens and his colleagues have found that simply noting which tests students choose provides only weak evidence about their thinking. Rather, it is sequences and, more specifically, ordered pairs of tests that are indicative of level of understanding. That is, knowledgeable students base their choice of subsequent tests on the results of the current test, in contrast to a trial-and-error or "do every test" approach characteristic of less knowledgeable students.

From a situative perspective, White and Frederiksen (1998, 2000) have developed curriculum and assessments to help middle school students acquire appropriate mental models for basic physical laws and their application across situations. For example, in Thinker Tools, computer-based representations are deployed to challenge students' existing conceptions of Newtonian models of force and motion. Cross-student debates and collaborative experimentation are used to resolve discrepancies between what students think and what the evidence from various inquiries or models seems to demonstrate. A cyclical sequence of "hypothesize, test, and generalize" is promoted and supported by the software and the overall instructional design. The goal is to support students' reflections on what they (individually and collectively) are doing and learning (i.e., metacognition) so as to promote the development of understanding. Opportunities for peer and self-assessment ("reflective assessment" in White and Frederiksen's terms) are an integral part of the teaching, learning, assessment cycle.

As these examples demonstrate, technology can extend the nature of the problems that can be presented and the kinds of knowledge and processes that can be elicited as evidence of student knowing. Innovation and utility notwithstanding, ongoing efforts to harness the potential of technology to support cognitively grounded assessments have been constrained by the high cost of "from-the-ground-up" development and lack of sufficient resources to keep pace with continuous technological advances (particularly the Internet). Further, technology-supported assessments, especially those designed for use in specific instructional environments, have been criticized for their limited applicability. These criticisms arose in part because the assessments could not be adapted for large scale use and in part because they were not well suited to adaptation or implementation outside the specialized context in which they were developed (Means & Haertel, 2002). In recent years, a number of industry-wide efforts have arisen to address these concerns and to meet the instruction and assessment development demands stemming from increased availability and use of technology in educational settings. Broadly speaking, these efforts seek to identify common elements and processes that could be programmed as objects (reusable parts) to support portability, platform independence, and long-term usability.

Two ongoing efforts to develop interoperability standards are noted here. The first, Shareable Content Object Reference Model (SCORM) (Randall House Associates, n.d.), is an XML-based framework used to define and access information in ways that permit it to be shared across various learning management systems (LMS). SCORM facilitates moving course content and related information (such as student records) from one platform to another, making course content into modular objects that can be reused in other courses and enabling any LMS to search others for usable course content. The second, IMS Global Learning Consortium, Inc. (n.d.), is developing and promoting open specifications for facilitating online distributed learning activities such as locating and using educational content, tracking learner progress, reporting learner performance, and exchanging student records between administrative systems. As part of this effort, IMS Question and Test Interoperability (QTI) standards specify protocols for exchanging assessment information such as questions, tests, and results. IMS/QTI standards are extendable and can be augmented to accommodate, for example, interactive computer- and Web-based tasks.

Common to IMS and SCORM is an effort to develop standards for software design to enable components of the programs to be reused or repurposed regardless of the particular technology environment. Reusability is accomplished in part by the use of objects—code-based abstractions of real-world entities or relationships. Objects consist of data and a set of behaviors and constitute the building blocks of object models. An object model is a group of related objects that work in concert to complete a set of related tasks. The PADI project applies the concept of object models to assessment design to facilitate generating, sharing, and reusing particular elements of the design process. As described below, the full PADI object model consists of structures including *design patterns*, *task templates*, and *task specifications* that lay out the elements of assessment design and the relationships among them. To support a broad range of designers (e.g., researchers, classroom teachers, commercial test publishers) and the corresponding variation in assessment tasks and uses, PADI objects can be extended, constrained, or wrapped within a user interface specifically suited to a particular purpose.

2.3 Measurement Methods and Techniques

A fundamental issue in measurement is summarizing and reporting on a set of performances in theoretically and empirically defensible ways; this in turn is bound up with the statistical representation of student performance. Too often, assessments simply indicate that some students have learned well, others not at all, and many are in between. Assessment practice has changed a great deal in response to evolving conceptions of knowledge and its acquisition, views of schooling and its purposes, and technologies for gathering and evaluating response data. The idea that we are drawing inferences about students from a limited set of observations has not changed. What has changed is the nature of the observations and what it means to know.

Increasingly common are situations in which multiple aspects of knowledge or skill are of interest. They are tapped in varying combinations by various tasks, and/or task performances provide several, often dependent, bits of information about various aspects of knowledge and skill. In these situations, probability-based models provide explicit, formal rules for integrating the many and diverse pieces of information that may be relevant to a particular inference about what students know and can do. The objective in the statistical model is to express, in probabilistic terms, the ways in which certain aspects of performance depend on particular aspects of knowledge. The relevant aspects of a student's performance are synthesized as probability distributions of variables that represent the targeted aspects of the student's knowledge. Item response theory models and latent class models are familiar examples of this kind of reasoning. Recent work has produced a variety of extensions that deal with multiple aspects of knowledge, skill, and strategy as they are seen from a cognitive perspective (Pellegrino et al., 2001; Junker, 2000). Depending on the purpose of the assessment, the nature of the observations, and the kinds of inferences one wishes to make, a given model will be more or less appropriate.

Consider, for example, a system of embedded assessments designed to guide teaching and inform learning of the Issues, Evidence, and You (IEY) curriculum developed at the Lawrence Hall of Science (Roberts, Wilson, & Draney, 1997; Wilson & Sloane, 2000). These classroom-based assessments are used to evaluate student progress on five important dimensions of decision-making: designing and conducting investigations, evidence and tradeoffs, understanding concepts, communicating scientific information, and group interaction. Over the course of the year-long curriculum, students are challenged to make decisions on a number of issue-oriented topics, such as "Water Usage and Safety" or "Environmental Impact." Assessments are administered within and between topics. Each assessment task is designed to measure student performance on one or more of the dimensions listed above. Although each task provides evidence for one or more (but not necessarily all) of the five key dimensions, student performance (and progress) is "mapped" in terms of the multiple dimensions the curriculum was designed to promote (Wilson, & Draney, 1997; Wilson, Draney, & Kennedy, 2001)

One approach to dealing with proficiencies that have many aspects is to model the variation in students and tasks at some level with multivariate models (e.g., Adams, Wilson, & Wang, 1997). From a multivariate perspective, each student can be characterized by more than one variable, each reflecting a distinct aspect of proficiency, and each task can be characterized by the degree to which it tends to stress the different aspects of proficiency. Now student-by-task interactions that render different tasks easy for some students and hard for others can be modeled and expressed as profiles of proficiency that differ among students. In contrast, the more familiar univariate approach simply characterizes each student by a propensity to do well on tasks from some specified domain; student-by-task interaction is viewed as measurement error. Thus, a multivariate approach allows for interpretation of student responses to complex problems in real-world situations and addresses the generalizability problem common to performance assessments (Linn, 1994; Shavelson, Baxter, & Gao, 1993).

In assessment situations that are cognitively motivated and supported by technology, Bayesian inference networks (Bayes nets for short) have proven to be broadly applicable in domains as diverse as electronics (e.g., Mislevy & Gitomer, 1996), dental hygiene (Mislevy et al., 2002), and physics (e.g., Martin & VanLehn, 1995). Bayes nets are representations of the probabilistic relationships among a set of variables (e.g., Almond, 1995; Pearl, 1988) that exploit conditional independence relationships to make inference feasible in even large networks of variables.[3] In educational assessments, attention focuses on the interrelationships between two kinds of variables: those concerning targeted aspects of knowledge and skill and those concerning observed performance. Bayes' theorem provides a mathematical expression of the probability that a student has the targeted knowledge/skill, given what we observe him/her do in an assessment situation. The power of this approach stems from the appropriation of prior information of the interrelationships between variables (from theory, expert judgment, or experience) to make predictions about (i.e., draw inferences from) the current situation from tasks constructed to best reveal those relationships.

VanLehn and his colleagues (e.g., Martin & VanLehn, 1995) use Bayes nets to evaluate what students know about Newtonian mechanics and kinematics. The online assessment of expertise (OLAE) collects data from students solving problems in introductory college physics and analyzes the data with probabilistic methods to determine what knowledge the students are using. Using an expert model, OLAE automatically creates a Bayes net that relates knowledge, represented as a set of rules, to particular actions taken during problem solving, such as equation writing. Having constructed a Bayesian network, OLAE can now "observe" a student's problem-solving behavior and compute the probability that the student knows and uses each of the rules. The focus is on what students know and the ways in which they use that knowledge, as opposed to a more traditional focus on how much students know (e.g., number of correct responses).

In each of these examples, the characterization of student knowledge/understanding relies on the interplay of substantive issues and psychometric/statistical technique. As definitions of what it means to know have changed, so too have the goals of schooling and the requirements of assessments. Consequently, familiar measurement models have evolved (and new ones have been developed) to make it possible to reason from assessment data to inferences about student achievement in an ever broadening range of situations (Junker, 2000). For example, "it is now possible to characterize students in terms of multiple aspects of proficiency, rather than a single score; chart students' progress over time, instead of simply measuring performance at a particular point in time; deal with multiple paths or alternative methods of valued performance; model, monitor, and improve judgments on the basis of informed evaluations; model performance at the level of students and also at the levels of groups, classes, schools, and states" (Pellegrino, Chudowsky, & Glaser, 2001, p. 168).

Despite these capabilities and the availability of computers to handle the computational requirements, these and other models and methods are not widely used. Although some are available in off-the-shelf packages, their use requires specialized knowledge. Much

[3] The interested reader is referred to Malakoff (1999) for a readable treatise and examples that extend beyond education.

work needs to be done to coordinate complex statistical models with current conceptions of knowledge and the kinds of performances indicative of more or less knowledge in a domain—a task that researchers are currently in a position to work out. *Knowing What Students Know* (Pellegrino et al., 2001) speculates that it will take time, as experience, examples, and tools accumulate, for less traditional psychometric methods to become more widely used in the science assessment community.

For its part, PADI includes formal probability-based reasoning, in the form of measurement models, as part of the evidence-centered design structure on which the PADI framework is predicated. In addition to knowledge representations such as *design patterns* and *task templates* for designing assessments, PADI is developing a "scoring engine" compatible with the PADI framework. The scoring engine is based on the work of Wilson and his colleagues with multivariate psychometric models (e.g., Adams, Wilson, & Wang, 1997) and includes submodels that deal with categorical, ordered, and conditionally dependent response variables (see below). As with *design patterns* and *task templates*, the scoring engine is presented as an extensible object model that can accommodate a family of models to meet the needs of various users and assessment uses.

3.0 PADI: A Framework for Assessing Science Inquiry

The Principled Assessment Designs for Inquiry (PADI) project is an NSF-sponsored collaboration among researchers and developers at SRI International, Lawrence Hall of Science, and the Universities of Maryland, Michigan, and California-Berkeley. The goal of the PADI project, broadly speaking, is to produce a conceptual framework and a collection of development resources for designing assessments of science inquiry, including but not limited to Web-based and performance tasks. More specifically, PADI is undertaking a special-case implementation of the evidence-centered assessment design (ECD) framework developed at Educational Testing Service by Mislevy, Steinberg, and Almond (2002). The ECD framework explicates the interrelationships among substantive arguments, assessment design elements, and operational processes without reference to particular content, purpose, or underlying cognitive theory. Rather, ECD provides a general approach and set of principles that are relevant for all types of assessment. PADI, in turn, provides general assessment-design data structures with exemplars specifically aimed at designing assessments of science inquiry.

3.1 Evidence-Centered Assessment Design

In designing and using assessments, the essential task is one of drawing inferences about what a student knows, can do, or has accomplished, from limited observations of what a student says or does. An evidentiary perspective focuses attention on the relationships among (a) what we want to infer about examinees (student model), (b) what kinds of situations enable us to evoke the necessary evidence (task model), and (c) how we can reason from observations in these particular situations to inferences about students more generally (evidence model). Student, task, and evidence models comprise the critical elements of an assessment argument.[4] Evidence-centered design defines these elements and the interrelationships among them and thus serves as a guide through the layers of interconnected decisions involved in developing a coherent assessment argument (see Table 1).

At the heart of ECD is the *conceptual assessment framework*, the stage at which the substantive, technical, and operational elements of the assessment argument are detailed. (See Mislevy, Steinberg, & Almond, 2002, for a detailed description.) Earlier phases/stages (i.e., *domain analysis*, *domain modeling*) to provide the substance for the assessment argument. Subsequent stages (compilation and delivery) fill in the technical details and carry out the processes that are necessary to maintain the integrity of the argument. (See Almond, Steinberg, & Mislevy, 2002, for a full description of a four-process architecture for assessment delivery systems.)

The stages or layers are generally sequential in that assessment design begins with stage I, *domain analysis*. However, stages may be (and often are) revisited during assessment design as information from one stage (e.g., assessment trials with students) suggests necessary changes to one or more of the other stages (e.g., what constitutes evidence).

[4] In *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001, p. 44) the terms *cognition*, *observation* and *interpretation* are used to describe the three essential elements of the assessment triangle.

Table 1. PADI Instantiation of General Principles and Stages of Evidence-Centered Design.

Evidence-Centered Assessment Design	Purpose/Description of Stage	PADI Framework for Assessing Science Inquiry
I. *Domain Analysis*	▪ Nature of knowledge, how people acquire it, how they use it. ▪ Definition of competence ▪ Development of competence/understanding ▪ Purpose of assessment	▪ Definition of Inquiry from standards documents ▪ Inquiry assessments used by curriculum developers and researchers ▪ Discussions with subject matter experts and review of literature on the development of inquiry
II. *Domain Modeling*	▪ Systematic structure for organizing information gathered in *domain analysis* stage. ▪ Narrative description of proficiencies of interest, ways of getting observations that evidence proficiency, and ways of arranging situations in which students provide evidence of targeted proficiencies.	*Design patterns*—narrative description of connections between inquiry standards and ways of obtaining evidence of what students know about inquiry. ▪ Pointers to other relevant information (e.g., exemplar tasks, other *design patterns*, reference materials). ▪ Content and grade independent.
IIIA. *Conceptual Assessment Framework* ▪ Student Model ▪ Task Model ▪ Evidence Model — Evaluation — Measurement	▪ Expression of targeted knowledge as variables ▪ Identification of features of eliciting situations as variables in task schemas ▪ Identification and summary of evidence: — Task level scoring — Summary scoring	*Templates*—detailed, technical description, blueprint, or specs for creating a family of tasks. ▪ Specifies student and task model variables, rules for evaluating performance (e.g., rubrics), psychometric measurement models.
IIIB. *Compilation* ▪ Task Creation ▪ Statistical Assembly ▪ Assessment Implementation	▪ Models for schema-based task authoring, ▪ Protocols for fitting and estimation of psychometric models, ▪ Strategies and algorithms for adaptive and nonadaptive test construction.	Outside the PADI project, with the exception of: ▪ *Exemplary tasks* produced by FOSS and BioKIDS partners in the PADI project ▪ Reference to the Berkeley Evaluation & Assessment Research Center's Item Calibration procedures for optional PADI scoring engine
IV. *Four-Process Delivery Architecture* ▪ Presentation ▪ Response Scoring ▪ Summary Scoring ▪ Activity Selection	▪ Data structures and processes for implementing assessments. ▪ Desire for interoperable processes and assessment objects	PADI object dodels promote design of assessment elements and processes to common IMS/SCORM standards. Optional PADI scoring engine available for users to incorporate in their assessment applications.

3.1.1 Stage I

At the *domain analysis* stage, the goal is to identify what is important for students to know, the situations in which one might observe evidence of knowing, the purpose of the assessment, and the constraints and contexts of the proposed use of the assessment. Information is compiled from cognitive psychology, subject matter standards, research in the disciplines, and other relevant sources of information on how and what students learn (e.g., curricular materials). Although this stage of assessment design is critical to sound assessment, PADI is not tasked with developing data structures or supporting tools for carrying out *domain analysis*. Rather, PADI structures are introduced at the next stage.

3.1.2 Stage II

At the *domain modeling* stage, assessment developers think through and lay out (in a nontechnical fashion) the elements of the assessment argument (i.e., student, task, and evidence models) using the information and resources compiled in stage I. In the PADI framework, this organization is facilitated by a *design pattern*. As described below, *design patterns* are guiding structures or schemas that describe the key elements of an assessment argument at a narrative rather than a technical level (Mislevy et al., 2003). Although the *design pattern* structure could be used to plan assessments in any content domain and from any psychological perspective, the instances being developed in PADI focus on science inquiry and stand on cognitive and sociocultural psychological bases.

3.1.3 Stage IIIA

At the *conceptual assessment framework* stage, the goal is to provide details (substantive, technical, and operational) for the assessment argument. In the PADI framework, the organization of the more technical details of the assessment argument is facilitated by *templates* (see below). *Templates* are used to detail specifications for families of tasks. *Task specs* in turn provide blueprints for individual tasks by selecting particular options expressed in the more general *templates*. Like *design patterns*, these structures—as structures—are applicable across content areas, assessment purposes, and psychological perspectives. As noted, PADI is focused on working through exemplars of science inquiry from a cognitive or sociocultural point of view.

3.1.4 Stage IIIB

At the *compilation* stage, the focus is on pulling together various elements of assessment development, such as task authoring, psychometric modeling, and assessment implementation. It is not within the scope of the PADI project to develop authoring systems to actually implement tasks. The intention, rather, is that the PADI conceptual framework and object model provide the infrastructure around which authoring systems could be tailored to the needs of a wide range of projects and users. FOSS and BioKIDS partners will develop and administer tasks as an essential part of evaluating the PADI framework.

3.1.5 Stage IV

During the final stage, *four-process delivery architecture*, the goal is to orchestrate the operational processes of an assessment (Almond, Steinberg, & Mislevy, 2002). With the exception of the optional scoring engine, PADI is not developing delivery system capabilities. As with authoring systems, the particulars of delivery systems can vary greatly from one assessment to another, especially with regard to purposes (e.g., diagnostic, large-scale) and platforms (e.g., paper-and-pencil, Web-based). Nevertheless, the shared conception, representational forms, object definitions, and IMS/QTI- and SCORM-compatible protocols enhance the efficiency of delivery system design by providing a common infrastructure that can support tailored implementation.

In summary, PADI applies the principles and structures of evidence-centered design to support the creation of high-quality assessments of science inquiry. Web-based structures including *design patterns*, *task templates*, and *task specifications* (in the form of an extensible object model) serve to guide developers through the interrelated decisions prerequisite to the development of a coherent assessment argument. In what follows, we elaborate on our initial work with *design patterns* and include brief comments about our current work with *task templates*, object modeling, the development of a scoring engine, and the design of exemplar tasks.

3.2 Design Patterns

Patterns and pattern languages are ways to articulate best practices, describe good designs, and capture experience in ways that make it possible for others to reuse this experience (Gardner et al., 1998). These patterns and pattern languages are used in diverse design fields, such as architecture (e.g., Alexander, Ishikawa, & Silverstein, 1977) and computer programming (e.g., Gamma, Helm, Johnson, & Vlissides, 1994) because of their explanatory power and generative utility. In PADI, we adopt the term *design pattern* to describe organizing schemas built on the principles of evidence-centered assessment design. An assessment *design pattern* assembles, in nontechnical terms, the elements of an evidence-centered assessment argument. By capturing the key relationships in the substantive domain in a way that presages the more technical design elements (i.e., student, task, evidence models), a *design pattern* provides a bridge between the content expertise and the measurement expertise needed to create an operational assessment. Although the structure of *design patterns* described below can be applied to assessment arguments in any domain, it will be in keeping with PADI's focus to develop the ideas in the context of science inquiry.

3.2.1 Defining Inquiry

The *design patterns* being developed as exemplars in PADI are intended to guide the design of assessments of science inquiry. The AAAS's (1993) *Benchmarks for Science Literacy* and the National Research Council's (1996) *National Science Education Standards* view inquiry as central to science and to the process of acquiring deep understanding of science content. Despite the shared emphasis on inquiry, the *Standards* and *Benchmarks* conceptualize inquiry in slightly different ways. The *Benchmarks* call attention to inquiry concepts that students at various grade levels should understand, while the *Standards* explicate abilities as well as "understandings." For example, the *Benchmarks* stipulate that by the end of eighth grade students should "know that if more than one variable changes at the same time in an experiment, the outcome of the experiment may not be clearly attributable to any one of the variables" (p. 12). In contrast, the *Standards* state that

"Students should develop general abilities, such as . . . identifying and controlling variables" (p. 145).

Although PADI is motivated by these emerging understandings of the nature of inquiry, it is not an objective of the project to propose a singular or authoritative definition of the term. Rather, its goal is to provide structures for expressing assessment arguments (in terms of *design patterns*) and instantiating them in tasks (in terms of *templates*), a goal that should be achievable under any perspective. *Design patterns* and *task templates* are *structures* that support, but do not dictate, the *substance* of an assessment argument. The PADI design framework is therefore offered as an open system, in that researchers and assessment designers will be able to lay out assessment arguments and build assessment tasks in accordance with their own views of inquiry. By providing a common structural framework, PADI aims to facilitate sharing, comparison, and debate on ways to conceive and assess inquiry in science—helping the community wrestle with the meaning of inquiry, rather than attempting to resolve the issue. The structure of *design patterns* will help frame assessment arguments around the vision that emerges of the nature of inquiry and ensure compatible ways of assessing students' knowledge/understanding of inquiry.

3.2.2 Design Pattern Attributes

Design patterns, like standards, cut across content areas. As a data structure, a *design pattern* contains *attributes* or constituent pieces of information that address the necessary elements of an assessment argument (Mislevy, 2003). Each *design pattern* details the knowledge or skill one wants to address, kinds of observations that can provide evidence about acquisition of this knowledge or skill, and features of task situations that allow the students to provide this evidence. In addition, each *design pattern* provides links to standards, other *design patterns*, *task templates* and *exemplary tasks* as appropriate. Table 2 provides a list of the attributes and a brief definition of each.

Table 2. Attributes of a PADI Assessment Design Pattern

Attribute	Definition
Title	A short name for referring to the design pattern.
Summary	Overview of relevant assessment situations and relation to targeted knowledge, skills, and abilities.
Rationale	Explains why this item is an important aspect of scientific inquiry and explicates the chain of reasoning connecting the inference of interest about student proficiency to potential observations and work products.
Focal knowledge, skills, and abilities	The primary knowledge/skills/abilities targeted by this design pattern.
Additional knowledge, skills, and abilities	Other knowledge/skills/abilities that may be required by this design pattern.
Potential observations	Some possible things one could see students doing that would yield evidence about the knowledge/skills/abilities.
Potential work products	Modes, like a written product or a spoken answer, in which students might produce evidence about knowledge/skills/abilities.
Potential rubrics	Some evaluation techniques that might apply.
Characteristic features	Aspects of assessment situations that are likely to evoke the desired evidence.
Variable features	Aspects of assessment situations that can be varied in order to shift difficulty or focus.
I am a kind of	Associations with other objects ("my parents") that are more abstract or more general than this object.
These are kinds of me	Associations with other objects ("my children") that are more concrete or more specialized than this object.
These are parts of me	Associations with other objects that contain or subsume this one. For example, a windshield is a part of an automobile.
Educational standards	Associations with (potentially shared) Educational standard objects.
Templates	Associations with (potentially shared) template objects.
Exemplar tasks	Associations with (potentially shared) task exemplar objects.
Online resources	Relevant items that can be found online (URLs).
References	Notes about relevant items, such as academic articles.

3.2.3 Examples of Design Patterns

As this is written, PADI has compiled approximately 40 *design patterns*.[5] We selected initial *design patterns* to develop in one of two ways. First, an analysis of standards documents provided definitions of inquiry and statements of what was important for students to know and do. We adopted a broad view of inquiry to include not only ways of doing science but also unifying concepts and processes (e.g., evidence, models, and explanation) and perspectives on how students learn (e.g., Bransford, Brown, & Cocking, 1999). Second, a review of existing assessments developed for curricular projects or research studies

[5] PADI has developed one possible set of *design patterns*. Starting from a subject-specific perspective may result in a different set of *design patterns*. Indeed, the PADI framework allows for the addition of other *design patterns*.

provided examples of ways in which situations could be arranged to elicit information about students' understanding of various aspects of inquiry. Special attention was given to those assessments that specified a cognitive or situative perspective in their articulation of what was important for students to know and what constituted evidence of knowing. There is no claim that the PADI *design patterns* constitute a definitive set, nor is that the intent. Rather, the purpose of these *design patterns* is to create a shared language for communicating insight and experience about assessment design problems and their solution. In this way, we can document and clarify our collective understanding of what constitutes high-quality assessment design (i.e., coherent assessment argument). Summaries of three *design patterns* follow. The full *design patterns* are shown as Figures A1–A3 in Appendix A.

Viewing Real-World Situations from a Scientific Perspective. A scientific perspective acknowledges certain principles and structures as valid for understanding, explaining, and predicting the world around us. This *design pattern* is one of 10 we "reverse-engineered" from a series of integrated investigation problems (Center for Technology in Learning, SRI International, n.d.) developed to accompany the GLOBE curriculum.[6] To assess ability to investigate real-world problems, students were asked to analyze and interpret GLOBE data sets, then communicate their findings and conclusions (Quellmalz, Hinojosa, & Rosenquist, 2001). We created *design patterns* from GLOBE to reflect the foci of different phases of a structured investigation (i.e., planning, conducting, analyzing, comparing, interpreting, and communicating).

For the *design pattern* highlighted here, the focus is on the ways in which students frame a problem (i.e., scientific, personal, social, or political). To assess students' propensities to approach situations from a scientific perspective, they might be asked to critique responses given by others, describe how to solve a problem, or identify reasonable next steps. As with all the *design patterns* we have developed so far, this *design pattern* is not content specific but can be adapted by adjusting the structure of the setting. For example, the Chi, Feltovich, and Glaser (1981) problem-sorting experiment targets thinking about situations from a scientific perspective, but with a different content area and a different form. In their study, expert physicists were observed to sort problems into categories based on fundamental relationships such as equilibrium, Newton's third law, or conservation of energy; novices sorted the same tasks on the basis of surface features, such as having to do with pulleys, springs, or inclined planes.

Model Elaboration. A primary goal of scientists is the development of explanatory models that can be used to explore the natural world. As consistent or conflicting data accumulate, these models are subject to elaboration or revision, respectively. In education settings, students even at a very young age construct models to account for their observations in mathematics and science (Lehrer & Schauble, 2000). However, research has shown that there are often discrepancies between student models and scientific models (e.g., diSessa, 1982), thus making this aspect of science inquiry an important target of assessment.

The model elaboration *design pattern* is one of a suite of model-based reasoning *design patterns* developed from James Stewart's studies of genetics problem solving (Stewart &

[6] GLOBE curriculum is available online at www.globe.gov

Hafner, 1994). Model-based reasoning can be assessed in and of itself or as part of a larger investigation for which using models, model elaboration, or model revision also are assessed. For model elaboration, the *design pattern* highlighted here, students are asked to solve problems in which the data do not conflict with their existing models. Problem solution involves combining or making additions to existing models by, for example, embedding a model in a larger system, adding more parts to the model, or incorporating additional information about a real-world situation into the schema the model represents. As with many of the PADI *design patterns*, the model elaboration *design pattern* can be applied to any content area and any grade level. Elementary students, for example, may be working with a simple model of magnetic attraction, while college students work with molecular models for the transmission of inherited characteristics. The essential processes of model-based reasoning remain, as appropriate to the content, the contexts, and the learners.

Reflective Assessment. White and Frederiksen's work on the inquiry cycle attends to the socioculturally motivated issue of helping kids learn the standards of good inquiry, externally at first, and then coming to internalize them. "By reflecting on the attributes of each activity and its function in constructing scientific theories, students grow to understand the nature of inquiry and the habits of thought that are involved" (White & Frederiksen, 2000, p. 334). For this *design pattern*, the focus is on the ways in which students think about what they are doing (i.e., metacognition)—in particular, how they apply the standards of evaluation to their own work, both while it is in progress and when they are done. Metacognitive skills such as this are not content or age specific—we would like students from elementary through postsecondary education to do this type of content-based thinking in contexts in which they find themselves. Further, metacognitive skills may be appropriately assessed in conjunction with other aspects of inquiry, such as using models or conducting investigations. In these situations, multiple *design patterns* can be used together to design a task or set of tasks that can reveal multiple aspects of inquiry-based reasoning.

The examples described here speak to the breadth, flexibility, and utility of *design patterns*. *Design patterns* can characterize assessment arguments for multiple aspects of inquiry and/or various psychological perspectives (breadth). Moreover, PADI *design patterns* are content independent and can be combined with other design pattern, or adapted for particular purposes (flexibility). Furthermore, they provide guidance in laying out the essential information necessary to create high-quality assessments, regardless of the purpose of the assessment, grade level, or content (utility). It is important to note that for each *design pattern*, consideration is given to the targeted aspects of inquiry and to the additional knowledge/skills/abilities that may be required. For example, students' familiarity with the particular content, level of content knowledge required, or familiarity with the task context can greatly affect performance and therefore what the assessor can learn about what students are apt to do in various situations. Ways in which tasks can be varied to increase or decrease demands for knowledge are noted in each *design pattern*. The designer of an assessment task should take these design decisions into account and construct tasks that will be informative, given (a) the purpose of the assessment, (b) the students who will be assessed, (c) what else is known about the test takers' backgrounds, and (d) the constraints and resources that will shape the assessment context.

In summary, the power of *design patterns* is twofold. First, by capturing thinking about important aspects of inquiry-based reasoning and paradigmatic strategies for assessing them, *design patterns* provide a starting point for designing inquiry tasks. This kind of assessment design support is increasingly helpful as the goals of assessment and the nature of the knowledge and skills to be assessed become more complex. The *design patterns* offer accumulated wisdom about considerations for assessment in these contexts. Second, enormous value is gained by being able to refer to tasks as instances of particular *design patterns*. Similarities in assessments that may look very different on the surface are highlighted when the substantive intent of the tasks and design decisions that were made to address the knowledge/skill in particular ways for particular contexts are made explicit. This documentation that can then be shared, adapted, or repurposed for various users and uses.

3.3 Current Work

3.3.1 Task Templates

As described above, *design patterns* lay out the assessment argument in narrative fashion and provide the prerequisite substantive information for later stages in the design process. The more technical details of the argument are added in Stage IIIA, the *conceptual assessment framework* (see Table 1). To guide the technical aspects of assessment design, PADI is creating *task templates*.[7]

Templates coordinate task design in two ways. First, at a technical level, the structure of a *template* helps assure coherence among the disparate elements and processes that operate during an assessment, such as simulation environments, evaluation rules, reporting displays, and psychometric models. Important here is the coordination of specialists from different fields (e.g., content specialists, psychometricians, programmers, interface designers, and automated scoring coders), whose work must come together for a coherent assessment. Second, at a conceptual level, the substantive argument (as expressed in *design patterns*) continually guides technical design decisions in light of the purpose the assessment is meant to serve. This is an example of the “layered” approach to the design of complex systems that is typical of architecture and engineering (e.g., Brand, 1994). The conceptual layer addressed in *design patterns* focuses on the structure and content of a coherent assessment argument, without getting into the structures and the details of implementation. *Templates* focus on the structure and the details of the “pieces of machinery” that are needed to implement an assessment, while keeping the argument they are meant to instantiate in the background. It clarifies thinking to make both layers explicit and to work between them in the design process.

In PADI, the *templates* distinguish the structure of assessment elements from their content. To date, much insight has been gained by mapping existing assessments (reverse-engineering) into this common structure, as we have done with GLOBE, FOSS, and BioKIDS. The real power of the PADI *templates*, however, will come from making it easier to generate new tasks, even new kinds of tasks, without having to rediscover the elements

[7] For some detailed examples of work completed to date, the reader is referred to Riconscente, Mislevy, Hamel, & PADI Research Group (2005).

and relationships that underlie coherent assessment arguments and their instantiations in various assessment applications.

3.3.2 Object Models

A primary goal of PADI is to address limitations or shortcomings of earlier efforts to design technology-supported and other forms of performance-based assessments (e.g., scalability, cost-effectiveness, and replicability). To this end, PADI uses extensible object models and IMS/SCORM-compatible protocols to create Web-based tools (guiding structures) to aid the designer in incorporating his/her purpose, psychological perspective, and so on, into the elements of evidence-centered design. PADI object models can be used "behind the screen" by designers who want to adopt the PADI guiding structures but embed them in interfaces and data forms customized to their own assessment needs.

The full PADI object model consists of structures including *design patterns*, *task templates*, and *task specifications* that lay out the elements of assessment design and the relationships among them. As described above, *design patterns* address assessment at a conceptual level. *Task templates* and *task specifications* are technical objects—in essence, blueprints for creating and assembling the elements of implemented tasks (e.g., stimulus materials, tools for the student, evaluation rules, and psychometric models) in formats that are consistent with IMS and SCORM protocols. Using the *template* structures makes it possible to create assessment elements and processes that can be reused in different applications. For any given assessment, instances of the objects can be created to follow the assessment argument (expressed in one or more *design patterns*) in whatever ways are needed to suit the purpose and environments of that particular assessment.

3.3.3 Scoring Engine

With respect to a scoring engine, PADI will provide a family of psychometric models for supporting inferences from observations. PADI will extend the IMS/QTI standards to accommodate more complex measurement models (multidimensionality; partial credit, rating scale, and dichotomous observations; item bundles to deal with conditional dependence). This aspect of the project draws on the work of Wilson and his colleagues with multidimensional random coefficients multinomial logit models, or MRCMLM (Adams, Wilson, & Wang, 1997). Assessment designers could take immediate advantage of using the PADI scoring engine, or they could develop alternative scoring engines or bypass probability-based inference entirely, as it suits their purposes.

3.3.4 Exemplar Tasks

We will work with the science education community to design tasks using the PADI framework. To date, filled-in examples of *design patterns* and *task templates* have been reverse-engineered from GLOBE, BioKIDS, FOSS, NAEP, and TIMSS. Although this exercise has proven useful for development, the real power of the framework comes from the ability to generate similar or new tasks from a set or subset of the information (and experience) used to design existing assessments. Creating specifications for new families of assessment tasks in these applications and then authoring and field testing the resulting tasks represents the next major stage in our work. Results will be catalogued in a digital library of working exemplars of assessment tasks and accompanying scoring systems.

4.0 Concluding Comments

The importance of inquiry is emphasized in standards documents and curricular materials, yet this is the one aspect of science teaching and learning that is least likely to be assessed adequately. An explicit conceptual framework and a collection of development resources to guide the design of high-quality assessments of science inquiry can speed the diffusion of improved assessment practices. In this report, we detailed PADI efforts to formulate a design framework for science inquiry. The framework consists of a set of guiding structures, both conceptual and Web-based, that lay out the essential elements of a coherent assessment argument and make explicit the layers of associated design decisions. The goal, in part, is to realize, in the design of science inquiry assessments, the full potential of developments in technology, measurement modeling, and our understanding of learning and knowing in science.

More specifically, the PADI framework advances an evidence-centered approach to assessment design to ensure quality and continuity in the design process. An evidence-centered approach begins with a clear articulation of what it means to know and do science inquiry. In this context, the application of measurement models and statistical methods is necessary to make sense of the variation and complexity of performances observed in testing situations. Technology plays a central role in enabling these efforts to succeed by providing a link between conceptual and statistical elements of the design process. Further, Web-based guiding structures expressed as extensible object models address issues of limited replicability, scalability, and cost-effectiveness characteristic of many previous efforts to design complex assessments in meaningful contexts.

As the project proceeds, PADI is committed to (a) implementing the assessment design framework in an open-system object model that can be adapted by others to suit their assessment needs and inquiry perspectives, (b) developing supporting software to create and work with *design patterns* and *templates*, and (c) providing an initial set of high-quality exemplars to highlight the elements of a coherent assessment argument. The framework and supporting tools move developers beyond thinking about individual assessment tasks to seeing instances of knowing or achievement that are similar across content areas or skill levels. This construct-centered approach draws attention to reusable schemas for obtaining evidence about what students know from what they do or say or otherwise produce in an assessment situation. Second, designing assessment products within the PADI framework helps to ensure that the way in which evidence is gathered and interpreted evokes the underlying knowledge and serves the intended purposes the assessment is meant to address. Third, the common design architecture facilitates coordination among the work of different specialists, such as content specialists, statisticians, task authors, delivery-process developers, and interface designers.

Initial applications of the ideas encompassed in the PADI framework may be labor intensive and time-consuming. Nevertheless, the import of the ideas for improving assessment will become clear from (a) the development of working examples and (b) the identification of reusable elements and pieces of infrastructure—conceptual as well as technical—that can be adapted for new projects. The gains may be most apparent in the development of technology-based assessment tasks, such as Web-based simulations. The same conceptual framework and design elements may prove equally valuable in making assessment arguments explicit for research projects, performance assessments, informal classroom evaluation, and tasks in large-scale, high-stakes assessments.

References

Achieve, Inc. (2002, November). *Staying on course. Standards-based reform in America's schools: Progress and prospects*. Retrieved September 1, 2004, from http://www.achieve.org/achieve.nsf/Publications?openform

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. (1995). *Graphical belief modeling*. London: Chapman and Hall.

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*. Retrieved September 1, 2004, from http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy: Project 2061*. New York: Oxford University Press.

Anderson, J. R. (2000). *Cognitive psychology and its implications* (3rd ed.). New York: W. H. Freeman and Company.

Barron, B., Vye, N. J., Zech, I., Schwartz, D., Bransford, J. D., Goldman, S. R., et al. (1995). Creating contexts for community-based problem solving: The Jasper Challenge Series. In C. Hedley, P. Antonacci, & M. Rabinowitz (Eds.), *Thinking and literacy: The mind at work* (pp. 47-71). Hillsdale, NJ: Erlbaum.

Barron, B., Schwartz, D. L., Vye, N., Moore, A., Petrosino, A., Zech, I., et al. (1998). Doing with understanding: Lessons from research on problem- and project-based learning. *The Journal of the Learning Sciences*, *7*, 271-311.

Baxter, G. P., & Glaser, R. (1998). The cognitive complexity of science performance assessments. *Educational Measurement: Issues and Practice*, *17*(3), 37-45.

Brand, S. (1994). *How buildings learn: What happens after they're built*. New York: Penguin Books.

Bransford, J. D., Brown, A., & Cocking, R. (1999). *How people learn: Brain, mind, experience and school*. Washington, DC: National Academy Press.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*, 32-42.

Bruer, J. T. (1993). *Schools for thought. A science of learning in the classroom*. Cambridge, MA: MIT Press.

Center for Technology in Learning, SRI International. (n.d.). *GLOBE assessment tools.* Retrieved September 1, 2004, from http://globeassessment.sri.com/index.html

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121-152.

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum.

Cognition and Technology Group at Vanderbilt. (1994). From visual word problems to learning communities: Changing conceptions of cognitive research. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 157-200). Cambridge, MA: MIT Press/Bradford Books.

Cognition and Technology Group at Vanderbilt. (1997). *The Jasper Project: Lessons in curriculum, instruction, assessment, and professional development*. Mahwah, NJ: Lawrence Erlbaum.

diSessa, A. (1982). Unlearning Aristotelian physics: A study of knowledge-based learning. *Cognitive Science*, *5*, 37-75.

Ericsson, K. A., & Smith, J. (Eds.) (1991). *Toward a general theory of expertise: Prospects and limits*. New York: Cambridge Press.

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

Gardner, K. M., Rush, A. R., Crist, M., Konitzer, R., & Teegarden, B. (1998). *Cognitive patterns: Problem-solving frameworks for object technology*. Cambridge, UK: Cambridge University Press.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-47). New York: Simon & Schuster Macmillan.

Greeno, J. G., Pearson, P. D., & Schoenfeld, A. H. (1996). *Implications for NAEP of research on learning and cognition*. Report of a study commissioned by the National Academy of Education Panel on the NAEP Trial State Assessment.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to teaching physics. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 51-74). Cambridge, MA: MIT Press/Bradford Books.

IMS Global Learning Consortium, Inc. (n.d.). *Open specifications for interoperable learning technology*. Retrieved September 1, 2004, from http://www.imsproject.org/index.cfm

Junker, B. (2000). *Some topics in nonparametric and parametric IRT, with some thoughts about the future*. Commissioned paper for the United States National Academy of Sciences. Retrieved September 1, 2004, from http://www.stat.cmu.edu/cmu-stats/tr/tr710/tr710.html

Koedinger, K. R. & Anderson, J. R. (1993). Effective use of intelligent software in high school math classrooms. In *Proceedings of the World Conference on Artificial Intelligence in Education* (pp. 241-248). Charlottesville, VA: Association for the Advancement of Computing in Education.

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (ed.), *Advances in instructional psychology. Educational design and cognitive science.* Mahwah, NJ: Lawrence Erlbaum.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, *23*(9), 4-14.

Malakoff, D. (1999, November 19). Bayes offers a "new" way to make sense of numbers. *Science*, *286*, 1460-1464.

Martin, J., & VanLehn, K. (1995). Student assessment using Bayesian nets. *International Journal of Human-Computer Studies*, *42*, 575-591.

Means, B., & Haertel, G. (2002). Technology supports for assessing science inquiry. In *Technology and assessment: Thinking ahead* (pp. 12-25). Washington, DC: National Academy Press.

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

Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law,Risk, and Probability, 2*, 237-258.

Mislevy, R. J., Hamel, L., Fried, R., Gaffney, T., Haertel, G., Hafter, A., et al. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report 1). Menlo Park, CA: SRI International.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction*, *5*, 253-282.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3-66.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). *Making sense of data from complex assessments* (CSE Technical Report 538). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, University of California School of Education and Information Science.

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Quellmalz, E., Hinojosa, T., & Rosenquist, A. (2001). *Design of student assessment tools for the Global Learning and Observations to Benefit the Environment (GLOBE) program*. Presentation at the annual GLOBE International Conference, Blaine, WA

Randall House Associates, Inc. (n.d.). *Sharable Courseware Object Reference Model (SCORM)*. Retrieved September 1, 2004, from http://www.rhassociates.com/scorm.htm

Riconscente, M. M., Mislevy, R. J., Hamel, L., & PADI Research Group (2005). *An introduction to PADI task templates* (PADI Technical Report 3). Menlo Park, CA: SRI International.

Roberts, L., Wilson, M., & Draney, K, (1997). *The SEPUP assessment system: An overview*. BEAR Report Series, SA-97-1. Berkeley: University of California, Berkeley.

Rumbaugh, J., Jacobson, I., & Booch, G. (1998). *The unified software development process*. Essex, UK: Addison-Wesley Longman Ltd.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215-232.

Stevens, R. H., Lopo, A. C., & Wang, P. (1996). Artificial neural networks can distinguish novice and expert strategies during complex problem solving. *Journal of the American Medical Informatics Association*, *3*, 131-138.

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 Publishing Company.

University of California, Los Angeles (n.d.). *Interactive Multimedia Exercises (IMMEX)*. Retrieved September 1, 2004, from http://www.immex.ucla.edu

Vendlinksi, T, & Stevens, R. (2002). The use of artificial neural nets to help evaluate student problem solving strategies. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Fourth International Conference of the Learning Sciences* (pp. 108-114). Mahwah, NJ: Erlbaum.

Wilson, M., & Draney, K. (1997). *Developing maps for student-progress in the SEPUP Assessment System* (BEAR report series, SA-97-2). Berkeley: University of California, Berkeley.

Wilson, M., Draney, K., & Kennedy, C. (2001). *GradeMap* [computer program]. Berkeley: BEAR Center, University of California, Berkeley.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*(2), 181-208.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, *16*, 3-118.

White, B. Y., & Frederiksen, J. R. (2000). Metacognitive facilitation: An approach to making scientific inquiry accessible to all. In J. Minstrell & E. van Zee (Eds.), *Teaching in the inquiry-based science classroom* (pp. 331-370). Washington, DC: American Association for the Advancement of Science.

APPENDIX A

Three Examples of PADI Design Patterns

Figure A-1: Viewing real-world situations from a scientific perspective

View real-world situations from a scientific perspective | Design Pattern 9 [View Tree | Export]

Title:	View real-world situations from a scientific perspective		
Summary	A student encounters a real-world situation that lends itself to being framed from a scientific perspective. Does the student act in a way consistent with having done so?	Viewing a situation from a scientific perspective can be contrasted with, for example, personal, political, social, or magical perspectives. This is a design pattern that is clearly appropriate for younger students. It is also appropriate for adults, once they are outside their areas of expertise.	
Focal Knowledge, Skills and Abilities	Knowledge and understanding of how to view real-world phenomena from a scientific perspective.		
Rationale	A scientific perspective says that there are principles and structures for understanding real-world phenomena, which are valid in all times and places, and through which we can understand, explain, and predict the world around us. There are systematic ways for proposing explanations, checking them, and communicating the results to others. Tasks motivated by this design pattern seek to determine whether students bring this perspective to a situation either in which they are placed or that they create.		
Additional Knowledge, Skills and Abilities	Particular scientific content or models.		Designer can structure setting so that knowledge of particular scientific content or models either is required or is minimized.
Potential observations	Critiquing responses offered by other students, either predetermined or as they arise naturally.		
	Explaining how to get started investigating the situation.		
	Identifying reasonable scientific next steps.		
	Posing a scientifically-answerable question.		Question should be relevant, realistic, and potentially addressable in light of the situation.
Potential work products	Diagram of the situation.		Looking for relevant features, especially if there are particular substance or knowledge representations the student should be employing.
	Identification, from given possibilities, of those that reflect a scientific perspective.		
	Verbal (oral or written) question, explanation of how to get started investigating the problem, etc		
Potential rubrics			
Characteristic features	Motivating question / problem / situation.		
	Sufficient background information provided so student could potentially address it from a scientific point of view.		Background information is especially important for 'drop in from the sky' assessments. In instructional or curricular setting, however, a task can presume background information because students are known to be familiar with it.
Variable features	Amount of prompting/cueing.		Less cueing gives better evidence about whether student is internally inclined to see situations from a scientific perspective; more cueing gives better evidence about whether student is able to proceed knowing that it is appropriate to think from a scientific perspective.
	Amount of substantive knowledge provided.		When substantive knowledge, such as models, formulas, knowledge representations, tools, or terminology, is required for an appropriate response, to what degree is it provided? Providing them reduces the load on the substantive KSAs. Not providing them means the response requires, conjunctively, the substantive KSA and the focal inquiry KSA.
	Degree of substantive knowledge involved.		'Content lean' vs 'content rich' in Baxter and Glaser's terms. Light content focuses evidence on inquiry perspective. Heavier content puts stress on knowledge of that content and calls for seeing situation in terms of models/principles. This confounds the inquiry and content KSAs, but makes it possible to get evidence about whether the student sees situations scientifically with respect to given content. [Note: connects with diSessa research-see references entry below.]

(continued)

Figure A-1: Viewing real-world situations from a scientific perspective, continued

I am a kind of	Scientific Reasoning. This design pattern concerns a scientific problem to solve or investigate. Do they effectively plan ...
These are kinds of me	Design and conduct a scientific investigation. Students are presented with a scientific problem to solve or investigate. Do they effectively plan a... Plan systematic solution strategies. Students are presented with an open-ended problem to investigate and must generate a plan for solvin...
These are parts of me	Conduct investigations. Students are presented with a scientific problem to solve or investigate and a solution strategy. Do...
Educational standards	NSES 8ASI1.1. Identify questions that can be answered through scientific investigations. Students should develop t... Unifying Concepts 1.2. Evidence, models, and explanation
Templates	GLOBE Inquiry Template. GLOBE inquiry investigations take a student through phases of an investigation, using data from the ...
Exemplar tasks	GLOBE Activities. Almost all of the GLOBE assessment tasks require students to write a short report summarizing their ...
Online resources	http://globeassessment...
References	diSessa, A. (1982). Unlearning Aristotelian physics: A study of knowledge-based learning. Cognitive Science, 5, 37-75. Physics students solve complicated mechanics problems in the classroom, but fall back on naive explanations when asked what will happen next with kids on playground equipment- even though exactly the same models apply.
I am a part of	

Figure A-2: Model elaboration

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

Account Settings
Logout
Edit Model

Model elaboration | Design Pattern 84

[View Tree | Export]

Title:	Model elaboration	
Summary	This design pattern concerns working with mappings and extensions of given scientific models.	A central element of scientific inquiry is reasoning with models. This DP focuses on model elaboration, as a perspective on assessment in inquiry and problem-solving.
Focal Knowledge, Skills and Abilities	- Establishing correspondence between real-world situation and entities in a given model - Finding links between similar models (ones that share objects, processes, or states) - Linking models to create a larger, more encompassing model - Within-model conceptual insights	
Rationale	Scientific models are abstracted schemas involving entities and relationships, meant to be useful across a range of particular circumstances. Correspondences can be established between them and real-world situations and other models. Students use, and gain, conceptual or procedural knowledge working with an existing model.	Students' work is bound by the concept of an existing model (or models) so their work includes an understanding the constraints of the problem. Even though model elaboration does not involve the invention of new objects, processes, or states, it does entail sophisticated thinking and is an analogue of much scientific activity. Even though model elaboration does not involve the invention of new objects, processes, or states, it does entail sophisticated thinking and is an analogue of much scientific activity.
Additional Knowledge, Skills and Abilities	Familiarity with task type (e.g., materials, protocols, expectations) Subject-area knowledge	
Potential observations	- Catenating models across levels (e.g., individual-level and species-level models in transmission genetics) - Determining the degree to which observations correspond with predictions. - Explanation of modifications, in terms of data/model anomalies - Identifying ways that a model does not match a situation (e.g., simplifying assumptions), and characterizing the implications. - Mapping out the corresponding elements between a real-world situation and a scientific model.	
Potential work products	- Correspondence mapping between elements or relationships of model and real-world situation - Correspondence mapping between elements or relationships of overlapping models - Elaborated model - Written/Oral Explanation of reasoning behind elaboration	
Potential rubrics		
Characteristic features	Real-world situation and one or more models appropriate to the situation, for which details of correspondence need to be fleshed out. Addresses correspondence between situation and models, and models with one one another.	
Variable features	Is problem context familiar? Model given to student(s), vs. model to elaborate produced by student(s) themselves Must experimental work or supporting research be caried out in order to ground the elaboration? Single model to elaborate, vs. establishing correspondence among models at different levels or with different focus?	
I am a kind of	Scientific Reasoning. This design pattern concerns a scientific problem to solve or investigate. Do they effectively plan ...	
These are kinds of me		
These are parts of me		

(continued)

Figure A-2: Model elaboration, continued

Educational standards	
Templates	
Exemplar tasks	
Online resources	
References	Biomass project http://www.education.u... Marshall, S.P. (1995). Schemas in problem solving. Cambridge: Cambridge University Press. NSES standards 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. White, B. Y., & Frederiksen, J. R. (1998). Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students. Cognition and Instruction, 16(1), 3-118.
I am a part of	

Figure A-3: Reflective assessment

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

Hello lpecktheis
Account Settings
Logout
Edit Model

Reflective Assessment | Design Pattern 90 [View Tree | Export]

Title:	Reflective Assessment	
Summary	In this design pattern students are introduced to a process in which they learn to evaluate and assess their own and each others' research methods.	
Focal Knowledge, Skills and Abilities	Diagnose particular strengths and weaknesses	Reflective assessment makes students aware of the strengths and weaknesses of their current system or model. Self-evaluation encourages continual change and improvement, thereby discouraging unexamined models and ideas.
	Metacognitive skills	Learning to monitor the quality of one's thought and the product of one's effort. The implicit overall goal is teaching how to think about thinking. The metacognitive skills should compliment each other and be applicable to a wide range of cognitive contexts.
	Recognize the progress being made toward these objectives	A critique of the process itself. Students can be given the means to understand how to do well in their performances.
	Understand instructional objectives	Reflecting on what they have learned raises new questions.
Rationale	Reflective self-assessment helps students to be able to develop simultaneously the ability to monitor and improve their own learning as well as acquire subject matter. Additionally, understanding the criteria by which their work will be evaluated enables students to better understand the characteristics of good performance.	Reflective assessment directs learning as students begin to think more carefully about the qualities to strive for in a performance or product.
Additional Knowledge, Skills and Abilities	Communication and collaboration	May or may not be required, depending on whether the designer wants to encompass collaborative activity around reflective assessment.
	Self-awareness	Often the simple task of rating oneself can lead to reflection about what one really knows or can do and what areas are in need of improvement or better understanding.
	Subject area knowledge	Some tasks may require a strong knowledge of the subject area, as understanding one?s performance in that domain may not be measurable outside of the metacognitive skills.
Potential observations	Applying generally stated qualities of a rubric to the specifics of own or group's work	i.e., being able to map one's own work into the framework of evaluation.
	Explanation of rationale of process.	i.e., student explaining what s/he is doing when assessing own or group's products or performance.
	Identification of next step in a thinking cycle.	
	Recognizing and resolving contradictions between one's own and a standard work product	
Potential work products	Critique of Audio or video recordings/ transcripts of own or group's work	Allows the student to record a sample of behavior for subsequent self-analysis, off-line of having to do it while doing the work. Can be used as a form of scaffolding.
	Critiquing a flawed experiment/project	i.e., practicing reflective-assessment skills with work other than one's own, as a precursor to evaluating one's own work
	Self-assessment questionnaires	Designed to be completed by the student to assess performance on a certain task.
	Student produced rubrics for self-evaluation	Asking students to develop the rubric will highlight that they understand the processes they are looking for.
Potential rubrics		
Characteristic features	Guidelines/rubrics/standards for judging work	These guidelines embody the standards of good work, so that learning them in conjunction with applying them to their own work adds a layer to students' understanding of the domain.
	Work to which the guidelines ought to be able to be applied	Typically one's own or group's work.

(continued)

Figure A-3: Reflective assessment, continued

Variable features	❶	Amount of substantive knowledge required	Some tasks may require a strong knowledge of the subject area as understanding one?s performance in that domain may not be measurable outside of the metacognitive skills
		Formality of assessment	Reflective assessment can be more or less formal or informal. To highlight certain behaviors a more formal method is required, although more informal reflection can be encouraged for nearly any task. A more informal assessment may involve a conversation with the student about what steps they took whereas a formal assessment could involve a questionnaire, presentation, etc.
		Formative vs. summative assessment	Some tasks may require a strong knowledge of the subject area as understanding one?s performance in that domain may not be measurable outside of the metacognitive skills
		Formative vs. summative assessment *Specificity of metacognitive skills to particular task *Amount of prompting/ cueing	Some tasks may have several stages, allowing students the opportunity for reflection and improvement. *Some skills, such as checking one?s work, are more general cognitive skills, as opposed to some subject areas that require less generalizable skills. *In the initial stages of self-reflection, students will need to be prompted to look for certain criteria in their own work. This scaffolding may be removed as students develop more metacognitive skills; at this point selecting the appropriate self-monitoring skill may be more important.
		Group vs. individual reflective assessment	Assessment can be a social process where students can see how multiple perspectives can be applied in viewing one's own and others' work. Starting off as group work can also help students to practice, model for others, and internalize habits of reflection.
I am a kind of	❶		
These are kinds of me	❶		
These are parts of me	❶	Modifying solution strategies based on external feedback, self-monitoring, and reflection. In this design pattern, students engage in self-monitoring, reflection, and apply external feedback ...	
Educational standards	❶		
Templates	❶		
Exemplar tasks	❶	FOSS Self-Assessment for Forces and Motion. This is a technology-supported self-assessment in which students complete science problems related t... Thinkertools Inquiry Project. Emphasizes peer and self-assessment; having students monitor the quality of their own and others' wo...	
Online resources	❶		
References	❶	White, B. Y., & Frederiksen, J. R. (1998). Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students. Cognition and Instruction, 16(1), 3-118.	
I am a part of	❶		

Sponsor
The National Science Foundation, Grant REC-0129331

Prime Grantee
SRI International. *Center for Technology in Learning*

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