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*
Meredith Ittner and Klaus Krause, *Technical Writers/Editors*
Lynne Peck Theis, *Documentation Designer*

PRINCIPLED ASSESSMENT DESIGNS FOR INQUIRY
TECHNICAL REPORT 10

Cognitive Predictions: BioKIDS Implementation of the PADI Assessment System

Prepared by:
Amelia Wenk Gotwals, University of Michigan
Nancy Butler Songer, University of Michigan

Acknowledgment
This material is based on work supported by the National Science Foundation under grant REC-0129331 (PADI Implementation Grant).
Disclaimer
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

FIGURES

TABLES

ABSTRACT

Assessment is a key topic in the high-stakes, standards-driven educational system that is present today. However, many current assessments are not good measures of student understanding because they are based on outdated theories of how students learn and test only fact-based knowledge. In science, with the call for inquiry-based teaching and learning, we need measures that can gather information not just about science content knowledge, but also about inquiry skills and how science content and inquiry skills interact in students' abilities to reason about complex scientific ideas. This report examines how, with PADI support structures, we developed an assessment system for the BioKIDS curricular program. The report specifically outlines the theories and beliefs of learning that underpin the system, describes tools that translate this cognitive framework into tasks that elicit observations of important student knowledge, and uses data to interpret whether the cognitive framework behind the suite of tasks is predictive of student knowledge.

1.0 Introduction

With pedagogy and learning in science shifting toward inquiry-based methods of teaching and learning (National Research Council, 1996), the types of knowledge that are valued have changed. Goals for student learning in science now include not only increasing content knowledge but also developing scientific inquiry abilities. In the past, assessments of student knowledge focused mainly on content; however, if assessments examine only students' content knowledge, then inquiry is devalued and is less likely to occur in the classroom (Schafer, 2002). Therefore, new assessment instruments in science must be developed to systematically address both content knowledge and inquiry skills.

In the past several years, there have been significant advances, both in theories of learning and in measurement science, that have affected the way assessments are created and scored (Pellegrino, Chudowsky, & Glaser, 2001). New assessments of science inquiry must take into account the advances in science teaching and learning and in measurement capabilities. The BioKIDS: Kids' Inquiry of Diverse Species project (<http://www.biokids.umich.edu/projects/biokids.html>) has teamed with the Principled Assessment Designs for Inquiry (PADI) project (<http://padi.sri.com/>) to develop a support structure for the creation, implementation, and evaluation of science inquiry assessment tasks. In this report, we look at the cognitive theory that is the basis of this assessment system, the methods we used to translate our cognitive theory into actual assessment tasks, and what the interpretation of observed student responses tells us about our assessment system. The main research questions we address are:

- What is the cognitive framework used by BioKIDS in its application of the PADI assessment design system?
- How is this cognitive framework translated into tasks that elicit observations of inquiry skills?
- What does the interpretation of the observed results tell us about the predictive and systematic nature of our assessment system for both students' inquiry skills and content knowledge?

2.0 BioKIDS and PADI

BioKIDS: Kids' Inquiry of Diverse Species is a project, funded by the Interagency Educational Research Initiative (IERI), whose goals include the study of the longitudinal development of students' content and inquiry knowledge acquisition as they participate in several inquiry-based curricular units. The initial curriculum is an 8-week unit on biodiversity in which particular inquiry thinking skills are fostered through a carefully scaffolded activity sequence (Huber, Songer, & Lee, 2003). In particular, the curriculum focuses on scaffolding students' development of scientific explanations using evidence. In order to gain the tools necessary to follow students' learning trajectories as they participate in the BioKIDS curricula, the BioKIDS project has joined the PADI team to create inquiry assessments. The PADI project is also IERI funded, and its main focus is the development of a conceptual framework and software support tools for the systematic design of assessment tasks to measure scientific inquiry. PADI combines developments in cognitive psychology, research on scientific inquiry, and advances in measurement theory and technology to formulate a structure that supports the design of inquiry assessments. PADI team members, representing expertise in assessment design, technology, science content, and psychometrics, contribute to the development of the PADI design system and, in doing so, establish common terminology and design guidelines. The PADI design system promotes explicit decisionmaking that links "the elements of the design to the processes that must be carried out in an operational assessment" (Mislevy, Almond, & Lukas, 2004, p. 5).

3.0 Nature of Assessment

All assessments are based in a conception or philosophy about how people learn and what tasks are most likely to elicit observations of knowledge and skills from students; they are premised on certain assumptions about how best to interpret evidence to make inferences (Mislevy, Almond, & Lukas, 2004). Many assessments that are being used today are created by using a combination of various prior (many would argue, outdated) theories of learning and methods of measurement (Pellegrino, 2001; Pellegrino et al., 2001). For example, many large-scale assessments are based on the behaviorist learning theory that supports dividing complex skills into smaller pieces of knowledge and testing each of these pieces separately, as well as teaching and assessing ideas in abstract rather than contextual situations (Black, 2003). Pellegrino (2001) says the following:

> The most common kinds of education tests do a reasonable job with certain limited functions of testing, such as measuring knowledge of basic facts and procedures and producing overall estimates of proficiency for parts of the curriculum. But both their strengths and limitations are a product of their adherence to theories of learning and measurement that are outmoded and fail to capture the breadth and richness of knowledge and competence. The limitations of these theories also compromise the usefulness of the assessments. (p. 4)

The current theory of learning that many subscribe to, and that the inquiry approach to science learning is based on, is the constructivist theory. Constructivism is built on the belief that learners need to be active participants in the creation of their own knowledge and that students will learn better if they possess a schema on which to build new understandings and link new concepts (Bransford, Brown, & Cocking, 2000; Driver, Guesne, & Tiberghisien, 1985; von Glasersfeld, 1998). The kinds of assessments that are based on constructivism are likely to be considerably different from those based on behaviorism. Assessments in line with constructivist theories of learning move away from focusing on separate component skills and discrete pieces of knowledge and move toward examining the more complex aspects of student achievement, such as reasoning demonstrated in an inquiry-based science classroom (Pellegrino et al., 2001).

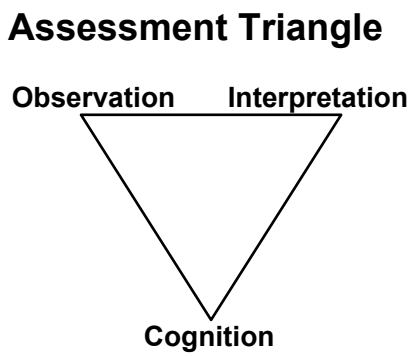
Assessment includes the processes of gathering evidence about a student's knowledge of and ability to use certain materials, as well as making inferences from that evidence about what students know or can do more generally for a variety of purposes (Mislevy, Wilson, Ercikan, & Chudowsky, 2002; National Research Council, 2001; Shavelson, Ruiz-Primo, Li, & Ayala, 2003). Assessment fulfills the "desire to reason from particular things students say, do, or make, to inferences about what they know or can do more broadly" (Mislevy, Almond, & Lukas, 2004, p. 6). One of the key steps in assessment is the actual design of tasks. In the past, task design was seen as more an art than a science; a more principled approach to assessment task design is needed, however, in order to be able to make the argument that a given set of tasks provide a good measure of the knowledge and skills being targeted (Mislevy, Steinberg, & Almond, 1998). Therefore, the design of complex assessments (like those needed to assess inquiry skills) must "start around the inferences one wants to make, the observations one needs to ground them, the situations that will

evoke those observations, and the chain of reasoning that connects them" (Messick, 1994, p. 20).

3.1 The Assessment Triangle

The National Research Council has illustrated the assessment process as an assessment triangle, the three corners of the triangle being cognition, observation, and interpretation (Pellegrino et al., 2001). In the assessment triangle, (1) *cognition* refers to the learning theory behind and the articulation of the knowledge that we are interested in measuring, (2) *observation* refers to the type of task that would best elicit performances that demonstrate an understanding of this knowledge, and (3) *interpretation* refers to a method of interpretation of the performance to make sense of the observations gathered from the task (Pellegrino et al., 2001). For a coherent and effective assessment, each corner of the triangle must not only make sense on its own, but must also connect to the other corners in clear and meaningful ways (Pellegrino et al., 2001).

Figure 1. NRC assessment triangle.



3.2 Evidence-Centered Design

Although the assessment triangle provides a good illustration of the nature of assessment, more elaboration is needed to develop a system for creating and evaluating assessment items. One such approach is the evidence-centered design (ECD) assessment framework (Mislevy, Almond, & Lukas, 2004). A quote from Messick (1994) functions as a grounding for understanding the principles underlying ECD:

> A construct-centered approach [to assessment design] would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 16)

The three main components of ECD that we focus on for our assessment design are Student Models, Evidence Models, and Task Models. Similar to the interconnectivity of the vertices of the assessment triangle (Pellegrino et al., 2001), each of these models must be explicitly linked to one another in order to create a functioning assessment argument

(Gotwals & Songer, 2004; Mislevy, Almond, & Lukas, 2004; Mislevy, Steinberg, Almond, Haertel, & Penuel, 2003; Pellegrino et al., 2001).

A Student Model defines the knowledge, skills, and abilities (KSAs) deemed important for the assessment at hand (Mislevy, Almond, & Lukas, 2004). For the case of BioKIDS, Student Models explicate the content knowledge and the complex reasoning skills the curriculum fosters. In the BioKIDS curriculum, the main content areas fostered are related to biodiversity (animal abundance and richness), animal interactions (food chains and food webs), and animal classification (defining animal characteristics). There are three main complex reasoning skills that the curriculum fosters through scaffolded activities: formulating scientific explanations from evidence, interpreting data, and making hypotheses and predictions. Because we know that inquiry in the classroom can take various forms and can occur at many different levels (Songer, Lee, & McDonald, 2003), our Student Models must take into account that students may be at many different places in their developmental range of being able to reason with complex scientific information and that it is important to recognize each of these levels. Through assessment tasks, these KSAs can be measured at different levels to examine students' development of knowledge and skills through participating in the curriculum.

Evidence Models are used to answer the question "What behaviors or performances would enable us to determine if students possess the knowledge, skills, and abilities defined in the Student Model?" (Mislevy, Steinberg, et al., 2003). In simple cases where we are curious only about students' definitional knowledge, having students offer a definition of a term or having them choose from a variety of given definitions may allow us to know whether students possess this type of knowledge. However, when examining students' complex reasoning, simple definitions are not adequate in determining students' abilities. Rather, having students do something such as creating explanations through making claims about scientific situations and backing up their claims with evidence is necessary to determine what type and level of knowledge or skills students possess. Because BioKIDS is interested in students' development of content knowledge and their acquisition of complex reasoning skills, students must be given the opportunity to demonstrate these types of performances in order to determine whether the goals of the curriculum have been met.

Finally, Task Models describe how to create and structure the kinds of situations and tasks needed in order to obtain the kinds of evidence required by the Evidence Models (Mislevy, Almond, & Lukas, 2004). In the case of discovering a student's declarative content knowledge, a simple multiple-choice type of question may suffice. However, gathering information about students' complex reasoning abilities, such as their ability to formulate a coherent scientific explanation, requires different types of tasks that can probe these deeper types of understandings. To deal with students' developing content and reasoning skills over the course of the curriculum, we needed items with a range of difficulties.

4.0 Domain Analysis

In the ECD framework, one of the first steps in assessment design is *domain analysis*. Similar to the cognition corner of the assessment triangle and the Student Models described above, *domain analysis* focuses on gathering important information about the domain of interest that holds implications for assessment. This information includes content, conceptual ideas, terminology, representational forms, and other knowledge and abilities that are associated with the domain (Mislevy & Riconscente, 2005). In addition, *domain analysis* can address the way in which people gain knowledge and the learning theories underlying the ways in which concepts are taught. The backbone of the BioKIDS project, and thus the main focus of our *domain analysis*, is scientific inquiry and the ways in which it is taught, learned, and measured.

4.1 The Importance of Inquiry

Before addressing inquiry-based assessment, it is important to understand the phenomenon of inquiry itself and why it is important in science education. The National Science Education Standards (National Research Council, 1996) state the following:

> Scientific inquiry refers to the diverse ways in which scientists study the natural world and propose explanations based on the evidence derived from their work. Inquiry also refers to the activities of students in which they develop knowledge and understanding of scientific ideas, as well as an understanding of how scientists study the natural world. (p. 23)

The process of inquiry is modeled on the scientist's method of discovery. This view represents science as a constructed set of theories and ideas based on the physical world, rather than as a collection of irrefutable, disconnected facts. It focuses on asking questions, exploring these questions, considering alternative explanations, and weighing evidence. Inquiry is important is because it can provide students with "real" science experiences, for example, experiences with many important features of science as practiced by professional scientists (Brown, Collins, & Duguid, 1989). The National Science Education Standards also state the following:

> 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. (p. 23)

This view of the classroom is different from what can be seen in traditional science classrooms—whether elementary, secondary, or postsecondary. A prevalent approach to science teaching emphasizes the end point of scientific investigations and the facts in textbooks (Lunetta, 1998). In these classrooms, learning consists of students' memorizing vocabulary, facts, and formulas; viewing demonstrations; and performing recipe laboratory exercises. Assessments often take the form of multiple-choice tests, which tend to emphasize general recall of knowledge over complex reasoning skills.

In contrast, inquiry learning emphasizes experiences with fundamental scientific phenomena through direct experience with materials; by consulting books, resources, and experts; and by debate among participants (National Research Council, 2000). Inquiry-based learning goals emphasize high expectations, including understanding that goes beyond simple recall of information. Students are expected to reason with scientific knowledge through activities such as formulating explanations, creating hypotheses, making predictions, and interpreting data. Various inquiry methods have been shown to encourage the inclusion of all students in science classrooms and to promote greater student achievement gains in both scientific content and inquiry knowledge (Krajcik et al., 1998; Mistler-Jackson & Songer, 2000; Songer et al., 2003; White & Frederiksen, 1998). In inquiry-based science programs, students do not just memorize scientific facts; they are exposed to the whats, hows, and whys of science. For these reasons and others, the National Science Education Standards state that "inquiry into authentic questions generated from student experiences is the central strategy for teaching science" (National Research Council, 1996, p. 31).

4.2 Scaffolding of Inquiry

Despite the exciting results shown by a number of research groups, several studies have found that students struggle with the complex reasoning needed in inquiry situations (Krajcik et al., 1998; Lee, 2003; Lee & Songer, 2003; White & Frederiksen, 1998). In particular, middle school students have difficulties with several aspects of inquiry, including asking questions and making decisions concerning how best to proceed within an extended inquiry and how to deal with data (Krajcik et al., 1998). Van den Berg, Katu, and Lunetta (1994) found that relatively open investigations alone were insufficient to enable students to construct complex and meaningful networks of concepts, and that other strategies and supports were needed. However, while students often struggle with the complex reasoning associated with inquiry when they are left to themselves, if provided with educational supports or scaffolds, they are able to work with complex scientific information and participate in inquiry activities (Metz, 2000). Educational scaffolds are structures that are placed strategically in the learning process to help students better understand confusing or unfamiliar topics. Written scaffolds can be implemented in many ways, including student notebooks with written scaffolds such as prompts, sentence starters, or hints about different aspects of inquiry (Lee, 2003).

Lee (2003) found that although scaffolds are meant to fade, fifth-grade students who had constant scaffolding of explanation building performed better than their peers who had fading scaffolds—suggesting that at this age, inquiry skills are still difficult enough that students need to have constant and consistent support in this aspect of inquiry. For many students, this will be their first foray into inquiry-based science learning. Because we expect that the development of complex reasoning takes time, we desired an assessment system that could assess beginning, intermediate, and complex levels of reasoning tasks (Songer & Wenk, 2003). We wanted to be able to see students' progression both through a single curricular unit and across curricular units and to determine their level of reasoning ability at each stage.

5.0 Domain Modeling

The purpose of *domain modeling* is to translate concepts and issues identified in the *domain analysis* stage into a coherent assessment argument that can be used to guide the development of assessment tasks. The process of *domain modeling* allows us to explicate the purpose of a given assessment and begin to outline how tasks might be developed in order to provide evidence of students' knowledge or abilities. In the case of BioKIDS, the purpose of the assessments is related to the goals of the curriculum, which are focused around development of biodiversity content knowledge and three focal inquiry skills (formulating scientific explanations from evidence, interpreting data, and making hypotheses and predictions).

5.1 Design Patterns for Inquiry

Science standards documents (National Research Council, 1996, 2000) outline aspects of inquiry that are important for students to learn, but they do not provide a cohesive guiding structure to assess these skills. The PADI team has developed structures that provide guidance in translating inquiry-based and constructivist-inspired standards and curricular learning goals into assessment tasks that reliably measure scientific inquiry skills. *Design patterns* are the structures under which all of PADI assessment task design falls (Mislevy, Chudowsky, et al., 2003). PADI *design patterns* provide a standardized way to represent an assessment argument so that tasks, developed for a particular assessment context, are guided by a consistent frame of reference.

Design patterns have been used in other disciplines for many years. The *design patterns* that have been created and used in other fields can provide good analogies of how *design patterns* will function in the case of assessment design. One example is that of Georges Polti's (1868/1977) *The Thirty-six Dramatic Situations*. Polti claimed that all literary works are based on and can be categorized into 36 dramatic situations, such as "falling prey to cruelty or misfortune" and "self-sacrifice for kindred." Polti explicated these situations or themes both to show similarities among dramatic stories and to provide a guide for authors in creating their own literary works. The dramatic situations are not meant to stifle writers' creativity or limit their creations; rather, these *design patterns* can be used as valuable resources for analyzing existing literature, as well as helping authors generate new stories (Mislevy, Chudowsky, et al., 2003).

The concept of *design patterns* also is present in fields such as architecture and computer programming (Alexander, Ishikawa, & Silverstein, 1977; Gamma, Helm, Johnson, & Vlissides, 1994), where, again, the *design patterns* are used both to analyze or classify preexisting artifacts (such as buildings or programs) and to provide structure for the creation of new works.

In all of these fields, the *design patterns* provide developers with tools to create new products. However, they do not give specific guidelines for any given story, building, or program. It is the developers' job to use their own creativity along with the scaffolds provided by the *design patterns* to construct new products. *Design patterns* for assessment are used to accomplish the same goals. They provide assessment developers a description or characterization of how a certain pattern of elements can be applied in several

situations, and developers in turn use *design patterns* to create new assessment tasks (Riconscente, Mislevy, & Hamel, 2005).

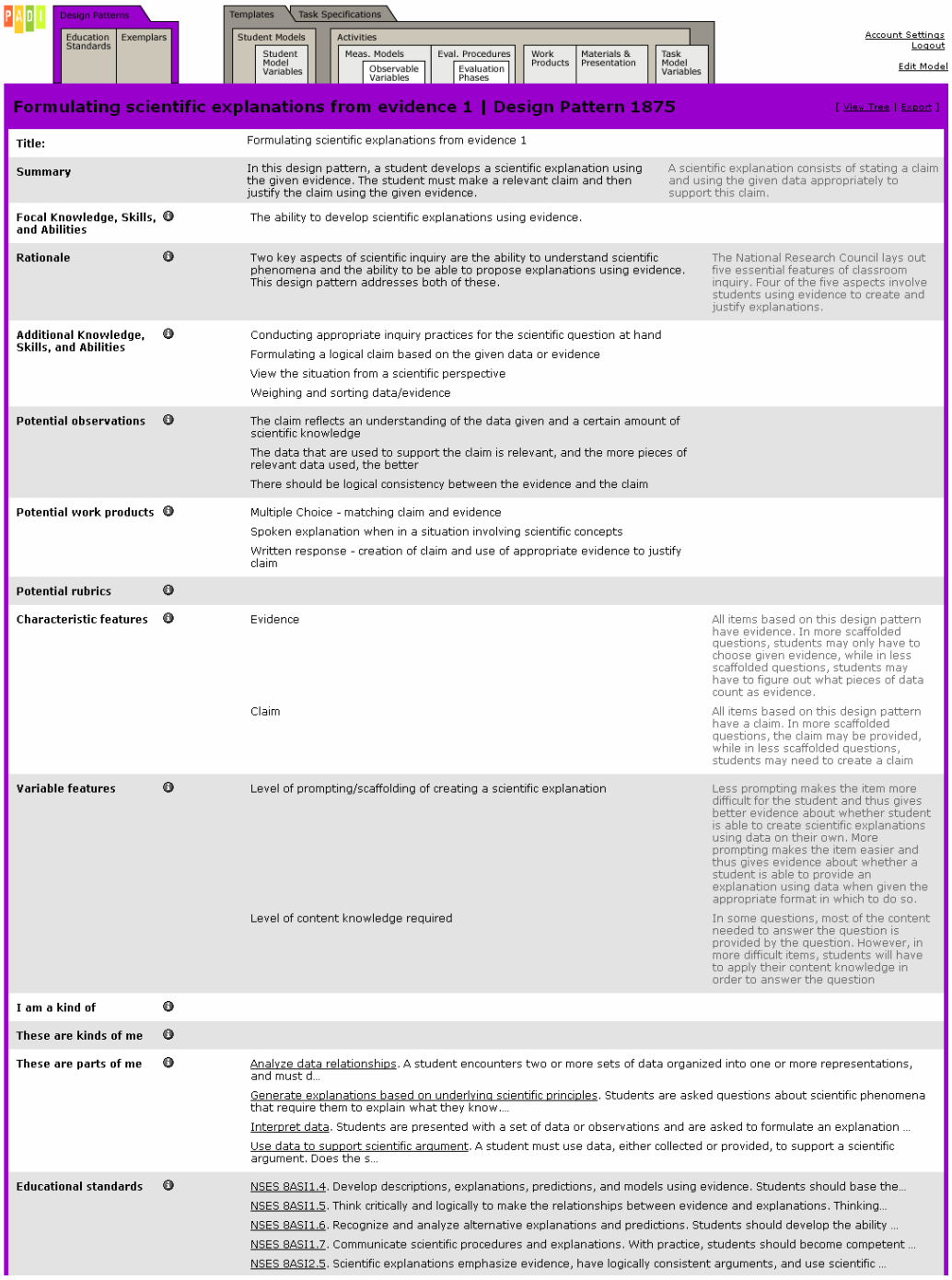
Specifically, PADI *design patterns* serve as a bridge between the science content and inquiry skills that are taught and learned in the classroom and the varying and complex ways in which they must be assessed in order to get an accurate account of what students know. They help to link assessment goals (consisting of content and inquiry standards and curricular learning objectives) with appropriate assessment task models and formats. To build this connection, *design patterns* outline "the chain of reasoning, from evidence to inference" by making explicit the three essential building blocks of an assessment argument: (1) the knowledge, skills, and abilities (KSAs) related to the aspect of inquiry to be assessed; (2) the kinds of observations one would like to see as evidence that a student possesses these KSAs; and (3) characteristics of tasks that would help students demonstrate these KSAs (Mislevy, Chudowsky, et al., 2003, p. 21). Specifying these features is the first step in creating an assessment task that can accurately measure some of the complex reasoning skills presented in inquiry-based science classrooms. "Making this structure explicit helps an assessment designer organize the issues that must be addressed in creating a new assessment" (Mislevy et al., 1998, p. 17).

In the BioKIDS project, we focus on three main *design patterns* related to the inquiry skills focused on in the curriculum: "formulating scientific explanations from evidence," "interpreting data," and "making hypotheses and predictions." To illustrate how *design patterns* can be used as a tool, we will focus on "formulating scientific explanations from evidence," one of the key aspects of inquiry that the BioKIDS curriculum fosters by using direct scaffolding. Figure 2 lays out attributes of the *design pattern* based on this aspect of inquiry. The first sections of the *design pattern* describe the aspect of inquiry being targeted (formulating explanations) and explain why it is an important part of inquiry. The skill of using evidence to create and justify explanations appears in all but one of the National Research Council's (2000) five essential features of classroom inquiry, making it an "essential essential." An explanation of the importance of the inquiry skill appears in the Rationale section of the *design pattern* table. In line with the assessment triangle and the ECD framework discussed above, the *design pattern* table provides space to list the Focal KSAs and Additional KSAs targeted by this aspect of inquiry. Clearly, the main skill in this *design pattern* involves the ability to formulate an explanation. However, being able to "formulate scientific explanations from evidence" could also involve other aspects of inquiry, like interpreting and analyzing data or the ability to view a given situation from a scientific perspective. These related skills, such as "interpreting data," are not necessarily used in all tasks that assess explanations, but they are closely related skills that may be used in tandem in certain assessment tasks.

The *design pattern* tool also provides room to articulate aspects of a task (Characteristic Features) that elicit the observations needed as evidence of the KSAs, as well as Potential Work Products that employ these features. For example, because in the BioKIDS project we define an explanation as consisting of a claim and use of evidence to back up the claim (Kuhn, 1989; Toulmin, 1958), observations we might look for would include confirmation that the claim represents an understanding of the given data and that students use appropriate and sufficient data to back up their claim. The kinds of tasks that we would

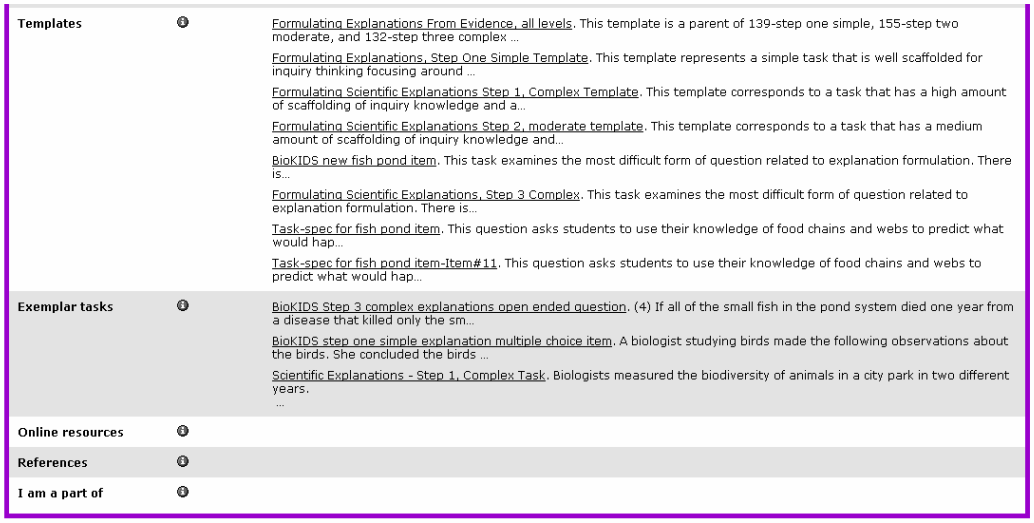
need to employ in order to gather information about students' ability to formulate explanations using evidence could take many forms. A simple question could use a multiple-choice format, while a harder question could ask a student to build an explanation without guidance. Although these tasks have different formats, they share certain Characteristic Features, such as including both a claim and evidence.

Figure 2. Design Pattern for "Formulating Scientific Explanations from Evidence"



(continued)

Figure 2. Design Pattern for "Formulating Scientific Explanations from Evidence" (continued)



5.2 Content-Inquiry Matrix

Although the tasks based on a single *design pattern* will have certain features in common, not all tasks associated with the same *design pattern* will be exactly alike. In fact, the ability to create a variety of tasks to address the same KSAs is one of the benefits of *design patterns* (Mislevy, Chudowsky, et al., 2003). Tasks stemming from the same *design pattern* can vary in terms of format, type of science content knowledge, and complexity. As inquiry in the classroom can take various forms and can occur at many different levels (Songer et al., 2003), it is important to develop tasks specifically oriented to different levels of complexity to accurately evaluate students' developing abilities over time. The Variable Features attribute of the *design pattern* table articulates some of the ways in which to vary the difficulty of the task.

In the BioKIDS project, we conceptualize science inquiry assessment tasks as having two dimensions of difficulty: the difficulty of the science content and the difficulty of the science inquiry. To address both aspects of task difficulty, we created a matrix that lays out three possible levels for each dimension, as shown in Table 1. First, we classified science content knowledge into three levels: *simple*, meaning that most content is provided by the task; *moderate*, meaning that students need a solid understanding of the underlying scientific concepts; and *complex*, meaning that students need not only an understanding of concepts but also the ability to link different concepts together.

Second, we separated inquiry into three levels: *step1*, *step 2*, and *step 3*. While the content aspect of the matrix can remain the same or similar for all *design patterns*, the steps of inquiry will be unique for each *design pattern* because of the inherently different nature of the aspects of inquiry being targeted. For the "interpreting data" *design pattern*, we specify the type of data that students are dealing with (table, graph, and so on); then, at each higher step, the interpretation of the data becomes more difficult, for example, through adding extraneous data into the item. For the "formulating scientific explanations from evidence" *design pattern*, we borrowed from our curricular units and created levels of

inquiry tasks based on the degree of support or scaffolding the task provides for forming explanations. *Step 1* tasks provide evidence and a claim, and students simply need to match the appropriate evidence to the claim (or vice versa). Although this only measures a low level of inquiry, specifically the ability to match relevant evidence to a claim (or a claim to given evidence), this is still an important step in students' development process. A *step 2* task involves a scaffold that provides students with a choice of claims and then prompts them to provide evidence to back up their choices. This involves more inquiry ability than the *step 1* task of matching, but there is still support for students to guide them in the important aspects of a scientific explanation. Finally, a *step 3* task is the most challenging in that it does not provide support in either the creation of a claim or the use of evidence. Students able to do *step 3* tasks demonstrate the knowledge of what is involved in a scientific explanation, as well as the ability and skill to construct such an explanation. We also have created similar matrices for the other two *design patterns* that we focus on: "interpreting data" and "making hypotheses and predictions."

Table 1. Levels of Content Knowledge and Inquiry Skill Needed for Assessment Items Related to the "Formulating Scientific Explanations from Evidence" Design Pattern

Level of Inquiry Skill Required for the Task	Level of Content Knowledge Required for the Task: **Simple** Minimal or no extra content knowledge is required, and evidence does not require interpretation.	**Moderate** Students must either interpret evidence or apply additional (not given) content knowledge.	**Complex** Students must apply extra content knowledge and interpret evidence.
Step 1 Students match relevant evidence to a given claim.	Students are given all of the evidence and the claim. Minimal or no extra content knowledge is required.	Students are given all of the evidence and the claim. However, to match the evidence to the claim, they must either interpret the evidence or apply extra content knowledge.	Students are given evidence and a claim; however, to match the evidence to the claim, they must interpret the data to apply additional content knowledge.
Step 2 Students choose a relevant claim and construct a simple explanation based on given evidence (construction is scaffolded).	Students are given evidence. To choose the claim and construct the explanation, minimal or no additional knowledge or interpretation of evidence is required.	Students are given evidence, but to choose a claim and construct the explanation, they must interpret the evidence and/or apply additional content knowledge.	Students are given evidence, but to choose a claim and construct the explanation, they must interpret the evidence and apply additional content knowledge.
Step 3 Students construct a claim and explanation that justifies the claim using relevant evidence (construction is unscaffolded).	Students must construct a claim and explanation; however, they need to bring minimal or no additional content knowledge to the task.	Students must construct a claim and explanation that require either interpretation or content knowledge.	Students must construct a claim and explanation that require them to interpret evidence and apply additional content knowledge.

In the past, classification of science performance assessments has been based on the amount of content involved and the freedom given to students in conducting scientific investigations. In particular, Baxter and Glaser (1998) identify four quadrants into which performance assessment tasks can be classified, based on the amount of content involved (content-rich or content-lean) and the amount of freedom students are given with regard to process or inquiry skills (constrained or open). Our matrix, shown in Table 1, looks at a similar dimension of content and a different dimension of inquiry. In addition to the level of content, we also found it important to look at the type of content knowledge required to answer the question. For example, some tasks require only understanding certain terms or groups of terms (like predator or prey), whereas other forms of content knowledge

require that students understand scientific phenomena (like disturbance of an ecosystem) and the interrelationships among these processes. Our matrix also examines the level of inquiry required to solve the task.

The main difference between our matrices and Baxter and Glaser's quadrants is that one of our matrices is specific to a single inquiry skill or *design pattern* (such as "formulating scientific explanations from evidence," "interpreting data," or "making hypotheses and predictions") and, in turn, outlines the Characteristic Features associated with each task in a given cell of the table. On the other hand, Baxter and Glaser's quadrants are created for scientific investigation performance assessments and are not specific to different inquiry skills (or *design patterns*). Instead, they group all skills involved in the investigation together. Both of these tools are useful in characterizing inquiry assessment tasks. Our matrices are more likely to be used to create tasks that measure specific inquiry abilities, whereas Baxter and Glaser's quadrants are more useful for designing and classifying performance assessments in which students conduct full scientific investigations.

6.0 Task Design and Mapping

Using the structure provided by the *design patterns*[1] and the content-inquiry matrices, we used both reverse and forward design processes to develop a coordinated set of assessment tasks measuring the three specific inquiry abilities in the BioKIDS curriculum. The reverse design process entailed mapping assessment items that had been used on past BioKIDS tests or on other assessments to existing *design patterns*, including mapping the level of content and inquiry involved. In addition to the explanations *design pattern*, BioKIDS assessment tasks also mapped onto multiple other *design patterns,* including "interpreting data," "reexpressing data," and "making hypotheses and predictions." Although some previously written items did fall into our matrices, we did not have a full set of assessment tasks at the end of the reverse design process. Therefore, we used the *design pattern* specifications, the matrix, and other structural components of the PADI system to forward design tasks. Developing new tasks occurred along a continuum of content and inquiry difficulty level associated with the focal biodiversity content and the three main aspects of inquiry (*design patterns*) that aligned with our particular curricular learning goals. In the end, we reverse engineered 7 tasks from previous BioKIDS tests, National Assessment of Educational Progress (NAEP) tests, and Michigan Educational Assessment Program (MEAP) tests, and forward engineered 9 tasks, for a total of 16 biodiversity tasks on our 2003 assessment.

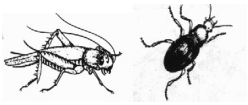
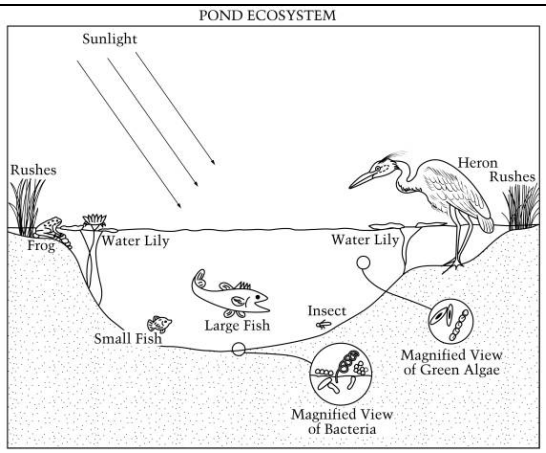
In mapping old tasks and creating new tasks, we used our content-inquiry matrices to make sure that we were examining all levels of content knowledge and inquiry skill. Most of the tasks fell into one of three categories: *step 1 simple*, *step 2 moderate,* and *step 3 complex*. We found that developing *step 1 simple*, *moderate*, or *complex* tasks was relatively easy; these tasks were generally multiple-choice questions with varying degrees of content difficulty. In contrast, we found it difficult to authentically address high levels of inquiry (mainly *step 3*) without involving content knowledge. This realization is congruent with our belief that inquiry skills are linked to content understandings, and that, particularly at higher inquiry levels, it may be difficult to tease apart content knowledge development from inquiry skills development. Thus, despite confounding inquiry skills and content understanding within our design, we focused on developing tasks along the diagonal of the matrices in our three *design patterns* (*step 1 simple*, *step 2 moderate*, and *step 3 complex*).

Table 2 provides examples of three tasks from the "formulating scientific explanations from evidence" *design pattern* that fall in the cells along the diagonal in the content-inquiry matrix shown in Table 1. As is clear from both the matrix and the examples, each level up (*simple* to *moderate* to *complex*) requires an increase in both the quantity and difficulty of the content knowledge. Our first question (*step 1 simple*) has a table that provides all the content information that a student needs to complete the task successfully; the student only needs to choose the relevant evidence from the table and match it to the provided claim. Our second question (*step 2 moderate*) provides students with pictures of invertebrates that they must group together, based on certain characteristics. Students are

[1] PADI has structures called *templates* that allow for the assessment argument to be more fully elaborated and more easily implemented. However, in our first try of designing and mapping tasks, the *templates* were not fully worked out. This report describes our first attempt at using the tools of *design patterns* along with the content-inquiry matrix. See PADI Technical Report 13 (Songer et al., in press) for an examination of the creation and use of *templates* for BioKIDS.

provided with the pictures so that they are not required to know all of the physical characteristics that separate insects from arachnids; however, to answer the question correctly, they do need to know what physical characteristics are important when classifying animals. Students are provided with a preformed claim statement in which they just have to choose the appropriate claim, and then they are prompted to give evidence. Finally, our last question (*step 3 complex*) gives students a picture, but this picture does not provide content information for the students. Students are provided a scenario, and they must construct (rather than choose) a claim and then, using their knowledge of food web interactions, provide evidence to back up their claim. While each of these questions targets students' ability to construct a scientific explanation from evidence, the tasks are clearly of different difficulty levels. Having these different levels is important if we want to measure students' developing inquiry skills. If we had only *step 1* questions (multiple-choice questions), we would not be able to see whether students could progress past the stage of matching claims and evidence. On the other end, if we had only *step 3* questions, we would not be able to determine whether students hold more tenuous skills for building explanations, which can only be evidenced with the presence of scaffolds. In addition, without a range of questions, we would not be able to accurately track students' development over time.

Table 2. BioKIDS Questions Mapped to the Level of the "Formulating Scientific Explanations from Evidence" Design Pattern

Question	Complexity Level
A biologist studying birds made the following observations about the birds. She concluded the birds would not compete for food. **Bird** **Food** **Feeding** **Where they feed** Bird 1 berries dawn/dusk trees, middle Bird 2 berries dawn/dusk trees, lower Bird 3 berries dawn/dusk trees, upper What evidence supports her conclusion? a. insects are plentiful b. they feed at different times c. they feed in different parts of the trees d. they lay eggs at different times	Step 1 simple
Shan and Niki collected four animals from their schoolyard. They divided the animals into Group A and Group B based on their appearance as shown below: Group A: Group B:   They want to place this fly  in either Group A or Group B. Where should this fly be placed? A fly should be in **Group A /Group B** Circle one Name two physical characteristics that you used when you decided to place the fly in this group: (a) (b)	Step 2 moderate
 ***http://nces.ed.gov/nationsreportcard/itmrls/*** 10. If all of the small fish in the pond system died one year from a disease that killed only the small fish, what would happen to the algae in the pond? Explain why you think so. 11. What would happen to the large fish? Explain why you think so.	Step 3 complex

7.0 Evaluating the Predictive Nature of the Assessment System

Despite having a solid cognitive theory that has guided task design through our *domain analysis* and *modeling* phases, no assessment can actually "get into" a student's head and measure exactly what he or she knows or can do. Thus, it is important to examine the nature of how students interact with the assessment tasks and compare it with how we hypothesized they would interact with the tasks.

7.1 Data Collection

To determine students' knowledge, skills, and abilities, we administered the BioKIDS test to a group of students and interpreted the results. In fall 2003, more than 2,000 sixth-grade students from 16 high-poverty urban schools participated in the BioKIDS curriculum. Twenty-three teachers with a range of experience and expertise taught the students. Students took both a pretest and a posttest made up of 16 questions, drawing from the diagonal cells of the matrices from each of the three focal inquiry skills (formulating scientific explanations from evidence, interpreting data, and making hypotheses and predictions). We used students' pretests and posttests to determine some properties of the assessment tasks.

7.2 Methods

7.2.1 Scoring

Although all the tasks for the BioKIDS assessments were created by using the same *design pattern* with specified levels of content knowledge and inquiry skill, we cannot equate any of the measures, psychometrically, as parallel or even as measuring the same construct. This is where the scoring and interpretation of scores come into play. On the basis of our cognitive model and previous student answers to questions, we developed a coding rubric to score all the tests. Multiple-choice items were scored 0 if they were incorrect or blank and 1 if they were correct. Open-ended items were scored differently, based on the question. For questions that addressed "formulating scientific explanations from evidence," we coded students' claim statement separately from their use of evidence to support their claim. Generally, the claim statement was scored 0 if it was incorrect or blank and 1 if it was correct. The score for the evidence portion was based on its scientific accuracy and the consistency between the claim statement and the evidence chosen, and a point was given for each relevant piece of evidence provided (up to two pieces of evidence). Table 3 shows the rubric for a *step 2 moderate* question. Because of the large number of tests, five people for whom greater than 90% interrater reliability was first established coded all the tests. For the rest of this report, we use these scores to examine some of the basic psychometric properties of the biodiversity assessment.

Table 3. Coding Rubric for Step 2 Moderate Item

Question	Point	Coding	Sample Responses
4. Shan and Niki collected four animals from their schoolyard. They divided the animals into Group A and Group B based on their appearance as shown below: Group A: Group B:  They want to place this fly  in either Group A or Group B. Where should this fly be placed? A fly should be in **Group A /Group B** Circle one Name two physical characteristics that you used when you decided to place the fly in this group: (a) (b)	1 2 Total = 3	(Claim) Correct (1) – Group A Incorrect (0) – Group B, multiple circles or no response (Data/Evidence) Complete (2) – two correct responses with no incorrect responses Partial (1) – one correct response; or two correct responses *with additional incorrect responses* Incomplete (0) – other responses or no response Correct responses include: ▪ having six legs/how many legs ▪ having wings ▪ having three body parts ▪ being insects ▪ not being spiders ▪ having antennae ▪ spiders and insects are not in the same group	**1** = The fly and the other insects have 6 legs and they have little eyes. They are called insects. Group B the spiders have 8 legs and they are animals. They have sharp nails/claws. **1** = Because the fly is an insect and so are the bugs in group A. Group B are both spiders. **0** = I chose group a because a spider is an animal not an insect. **0** = Because all of group A can fly. **0** = I put him there because of his legs.

7.2.2 Analysis

It is important to compare the mapped difficulty level of a question (*step 1*, *2*, or *3*) with the empirical difficulty level to determine whether the cognitive scheme guiding our task creation using *design patterns* and content-inquiry matrices maps with what students experienced when they completed the tasks. To determine the predictive ability and accuracy of our cognitive model, we used the student version of the Rasch modeling software Winsteps, which is called Ministep. We applied a Rasch (one-parameter) model, and although this model does not take into account discrimination of items or guessing parameters, it allows for a good estimate of difficulty level. Because Ministep has a limited capacity, SPSS software was used to randomly choose 100 students from our database who had completed both the pretest and the posttest. Choosing students randomly allows the results to be generalized to the whole database population. Because we have a mixture of multiple-choice items, which are coded as right or wrong (a binary code), as well as constructed-response questions, which are coded on a 0-1-2 scale, we had to run a partial-credit model that took the differing scales into account. To determine the relative difficulty of the items and to see how well matched they were to our population of students, we calculated the difficulty parameter and created item maps for both the pretest and the posttest.

Item response theory (IRT) models a student's response to a specific task or item in terms of an unobserved variable associated with each individual (McDonald, 1999; Mislevy et al., 2002). Each of these attributes (often termed latent traits or abilities) is posited to vary along a single dimension, usually denoted θ (Mislevy et al., 2002). From the perspective of trait psychology, θ may be thought of as an unobservable trait or ability. From the perspective of information processing, θ would be interpreted as a composite of the knowledge, skills, and abilities required to do well on tasks in the domain. From a sociocultural perspective, it is the strength of patterns of effective or ineffective action in the situations the students are learning to work in. BioKIDS combines the latter two of these perspectives, although for brevity and for continuity with the IRT literature we will refer to θ as "ability" below.

We used a unidimensional model to analyze these data. A unidimensional model assumes that there is a single latent trait or ability underlying how students respond to items. Because we only used items from the diagonal of our content-inquiry matrix, we can assume that both content and inquiry play a role in how students interact with the items.

The IRT ability continuum is set up as a standardized scale, with 0 being average and each number above and below 0 representing one standard deviation. Using IRT, assessment tasks and the students completing these tasks are placed on the θ scale from lowest to highest. The placement of student "i" on θ, (θ_i), is referred to as the student's ability or proficiency. The position of item "j" on θ (b_j) is referred to as the item's difficulty. In the item maps, items are lined up on the right-hand side of the divider, and one can determine the difficulty of the items by looking at their positions on the continuum (if they are close to 0, they have an average difficulty; above 0, they are more difficult and below 0, less difficult). On the left-hand side of the divider are Xs, which represent respondents. Respondents are arranged relative to their ability level. It is important to look at the relation of items and respondents on the θ continuum. Items are most informative for students whose ability level is closest to the item difficulty. For this analysis, we focus on difficulty level, so once the item maps were created, they were color-coded by matrix position regardless of *design pattern*.

7.3 Results

There are two difficulty tables and item maps, one for the pretest and one for the posttest. Tables 4 and 5 give a numerical difficulty value for each item. In the tables, bolded items are outliers, which are discussed in the following section. For the pretest, difficulties range from -3.32 at the least difficult to 2.64 at the most difficult. The range for the posttest is smaller, with the least difficult item having a difficulty of -2.50 and the most difficult item having a difficulty of only 2.11.

Table 4. Pretest Item Difficulty Listed from Most Difficult to Least Difficult

Item	Estimated Difficulty	Complexity Level
BioKIDS 14a	2.64	Step 3 complex
BioKIDS 13c	**2.26**	**Step 1 simple**
BioKIDS 13a	2.26	Step 2 moderate
BioKIDS 6d	2.03	Step 3 complex
BioKIDS 16a	1.95	Step 3 complex
BioKIDS 9	1.86	Step 2 moderate
BioKIDS 14b	1.68	Step 3 complex
BioKIDS 15	1.44	Step 2 moderate
BioKIDS 10	0.82	Step 3 complex
BioKIDS 8	0.71	Step 2 moderate
BioKIDS 13b	0.59	Step 2 moderate
BioKIDS 14c	0.47	Step 3 complex
BioKIDS 6e	0.42	Step 3 complex
BioKIDS 5a	0	Step 2 moderate
BioKIDS 4a	-0.08	Step 2 moderate
BioKIDS 11	**-0.69**	**Step 3 complex**
BioKIDS 2	-0.75	Step 1 simple
BioKIDS 6b	-0.92	Step 1 simple
BioKIDS 4b	-1.48	Step 2 moderate
BioKIDS 12	-1.48	Step 1 simple
BioKIDS 1	-1.76	Step 1 simple
BioKIDS 6c	-1.84	Step 1 simple
BioKIDS 7	-2.11	Step 2 moderate
BioKIDS 5b	-2.35	Step 2 moderate
BioKIDS 6a	-2.35	Step 1 simple
BioKIDS 3	-3.32	Step 1 simple

Note: Bolded items are outliers.

Table 5. Posttest Item Difficulty from Most Difficult to Least Difficult

Item	Estimated Difficulty	Complexity Level
BioKIDS 14a	2.11	Step 3 complex
BioKIDS 16a	2.04	Step 3 complex
BioKIDS 6d	1.76	Step 3 complex
BioKIDS 9	1.17	Step 2 moderate
BioKIDS 13a	1.16	Step 2 moderate
BioKIDS 14b	1.11	Step 3 complex
BioKIDS 15	1.11	Step 2 moderate
BioKIDS 13c	**1.05**	**Step 1 simple**
BioKIDS 14c	0.79	Step 3 complex
BioKIDS 8	0.55	Step 2 moderate
BioKIDS 10	0.54	Step 3 complex
BioKIDS 13b	0.48	Step 2 moderate
BioKIDS 5a	0.38	Step 2 moderate
BioKIDS 2	0.32	Step 1 simple
BioKIDS 4a	-0.03	Step 2 moderate
BioKIDS 6e	-0.12	Step 3 momplex
BioKIDS 6b	-0.18	Step 1 simple
BioKIDS 11	**-1.11**	**Step 3 complex**
BioKIDS 5b	-1.19	Step 2 moderate
BioKIDS 1	-1.36	Step 1 simple
BioKIDS 12	-1.36	Step 1 simple
BioKIDS 7	-1.38	Step 2 moderate
BioKIDS 4b	-1.56	Step 2 moderate
BioKIDS 6a	-1.77	Step 1 simple
BioKIDS 6c	-2.02	Step 1 simple
BioKIDS 3	-2.50	Step 1 simple

Note: Bolded items are outliers.

The item maps (see Figures 3 and 4 for the pretest and posttest, respectively) include the same information that is in the tables, only in a graphical structure with both the items and respondents on the same continuum (θ). Red items fall in the *step 1 simple* cell of the matrix, green items in the *step 2 moderate* cell of the matrix, and blue items in the *step 3 complex* cell. It is clear from the figures that the *step 1 simple* items (red) tend to be at the lower difficulty levels and the *step 3 complex* (blue) items tend to be at the higher difficulty levels, while the *step 2 moderate* questions have a broad range from low to high difficulty. However, there are a few exceptions that will be discussed in the next section.

In terms of item discrimination, for the pretest, items and student respondents tend to be generally aligned along the continuum; however, there are three students at a lower ability level (about -4.50) than we have questions for (the least difficult question is placed at -3.32

on the continuum). For the posttest, the opposite problem is true. While for the most part questions and students are matched, there are a few students who are at a high ability level, and we do not have questions matched to them. In addition, there is a gap in questions with difficulties -1.0 to 0, leaving slightly below-average students with no questions well matched to their ability level.

Figure 3. Item Map for Pretest, by Complexity Level ($N_{Students}$ = 100; N_{Items} = 26)

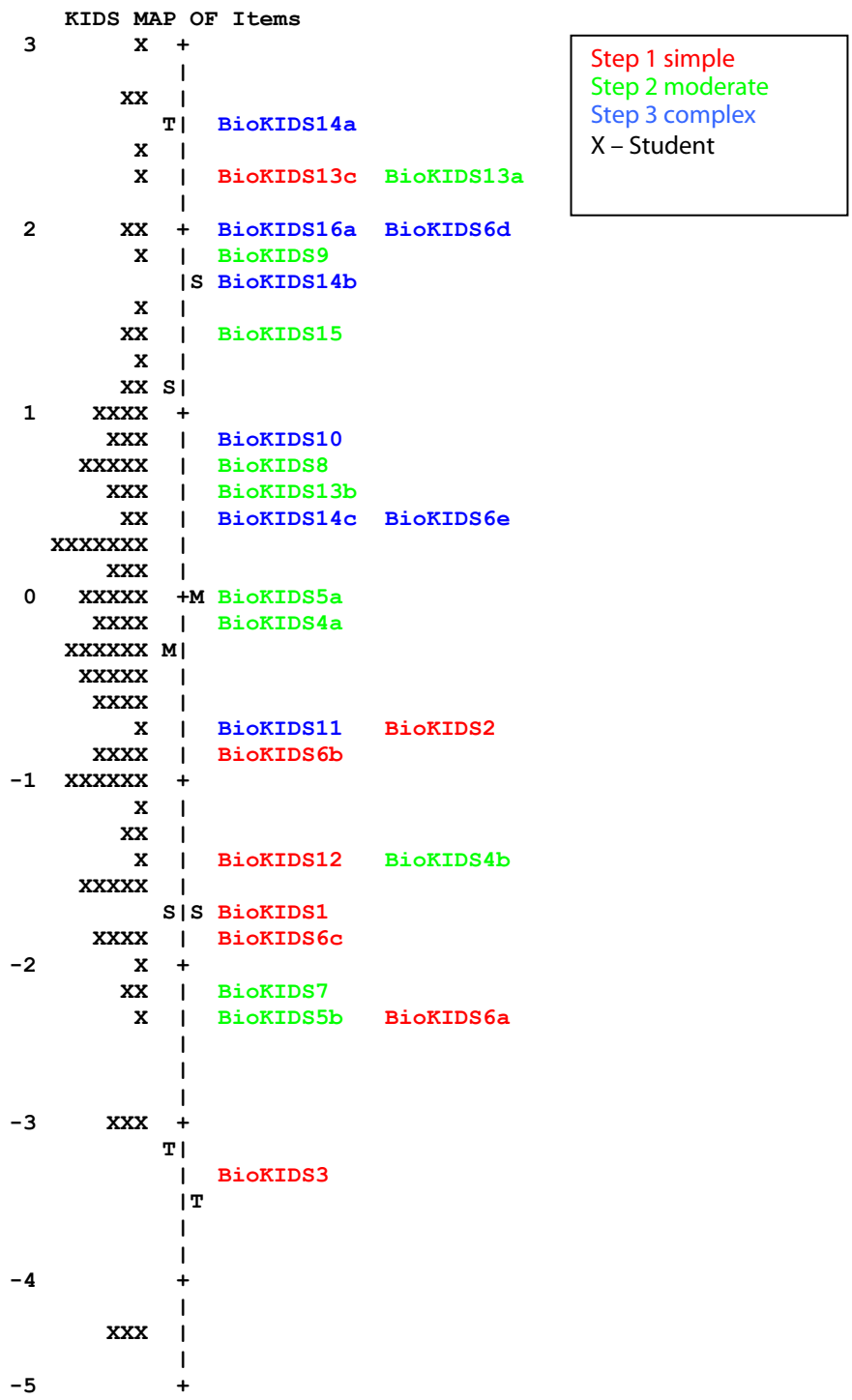
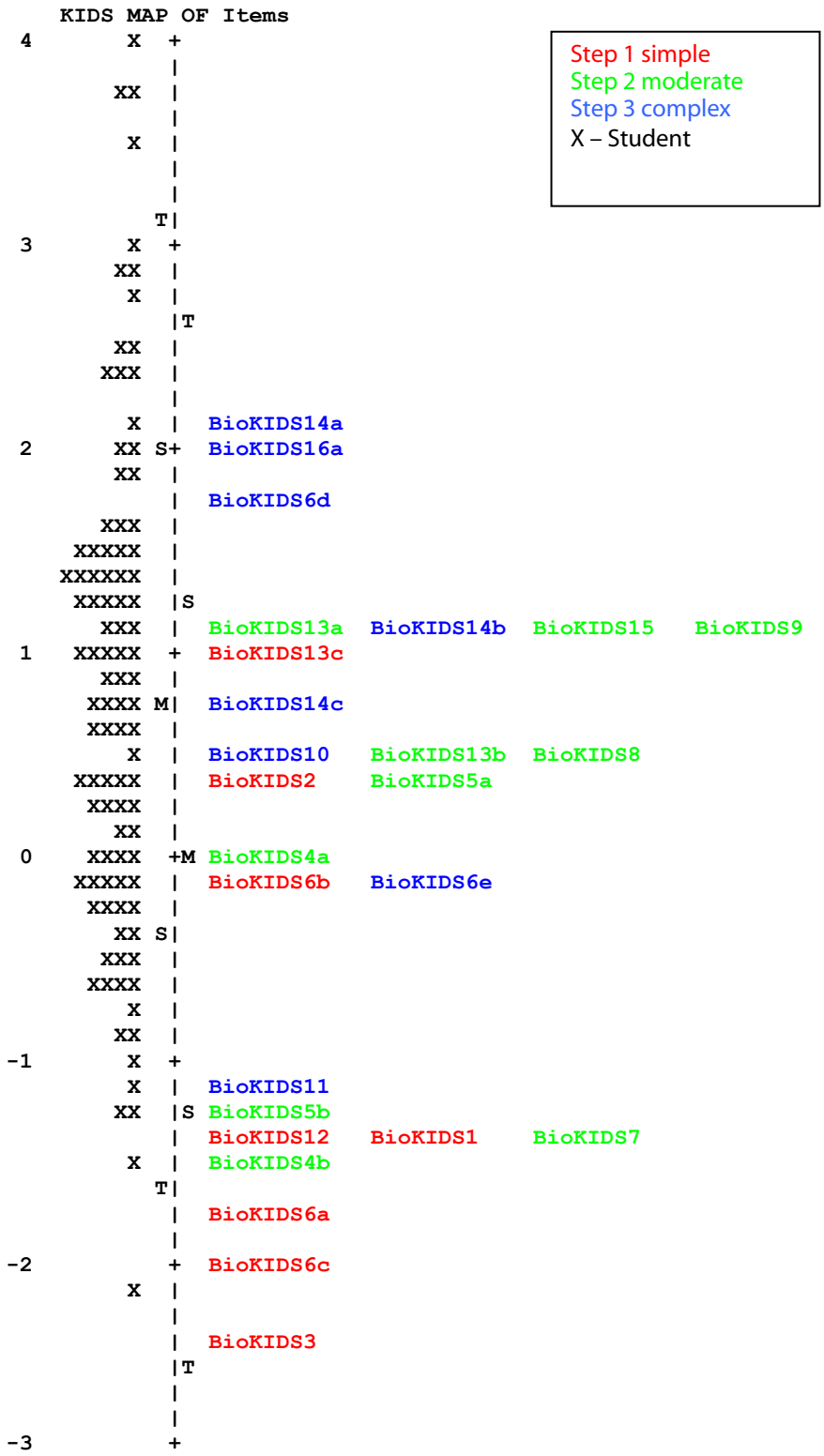


Figure 4. Item Map for Posttest, by Complexity Level ($N_{Students}$ = 100; N_{Items} = 26)



8.0 Discussion

The third research question posed was, "What does the interpretation of the observed results tell us about the predictive and systematic nature of our assessment system for both students' inquiry skills and content knowledge?" If our cognitive framework fit perfectly with the observed scores of students, we would expect to see all *step 3 complex* items with the highest difficulty, *step 2 moderate* items with a middle difficulty, and *step 1 simple* items with the lowest difficulty. Looking at the tables and the item maps, there are not these clean divisions between the groups of items. Generally, *step 3 complex* items have a higher difficulty and *step 1 simple* items have a lower difficulty, with the *step 2 moderate* items mostly in the middle section. However, there are "outliers" within each of two categories (*step 1 simple* and *step 3 complex* items). We now focus on these outliers and examine why they do not map well onto our cognitive framework.

BioKIDS item 13c is classified as a *step 1 simple* item, yet it is the second most difficult item on the pretest and has more than a standard deviation above average difficulty on the posttest. This item involves examining a table and determining which graph is the best representation of a column of that table (see Figure 5). In the BioKIDS program, we focus on three main *design patterns*: "formulating scientific explanations from evidence," "interpreting data," and "making hypotheses and predictions." This question fits into a separate *design pattern* called "reexpressing data." This question is the last of several items that are based on a single scenario. In this scenario, the other items are focused on "interpreting data" and "formulating scientific explanations from evidence." This section of the question asks students to recognize that the same data can be expressed in many forms—in this case, a table and a graph. Though this skill is addressed in the BioKIDS curriculum, it is not a main focus and is not directly scaffolded. Item 13c has students selecting a graph (not creating one), so it seems that, cognitively, it would be less difficult than other questions in which students have to construct their own responses. However, students' performance on this item indicated that they found the item to be difficult.

There are several factors that could make this item more difficult than we originally thought. Careful examination of this question shows several cognitive steps that students must go through to successfully answer this question. First, the labels on the graph axes ("Number of Animals" and "Type of Animal") do not match the labels in the table ("Abundance of Animals" and "Richness of Animals"), meaning that students need to know the definitions of "abundance" and "richness" in order to translate the data from the table to the data in the graphs. Second, students must take the numbers from the table and understand how to read a graph so that they can match the total number of animals (abundance) in the table to the total number of animals in the graph by adding the number of animals for each zone in the graph. In addition, the graph introduces names of animals that are not found anywhere in the table. Students need to recognize that they do not need to look for specific animal names; rather, they are looking for the total number of animal types in the zone (richness). However, having names on the graph may act as a distracter to students, causing them to focus on an unimportant aspect of the graph. Finally, another reason for the high level of difficulty of this item may be that reexpressing data is simply a difficult skill for sixth-grade students, and more scaffolding may be needed to make such items simpler for them to solve. Even labeling the axes of the graphs with the

terms “abundance” and “richness” might help students make the transition from the table to the graph. One positive aspect of this question is that the estimated difficulty of this item went from 2.26 on the pretest to 1.05 on the posttest, meaning that students learned some of the skills needed to solve this type of problem from the curriculum.

Figure 5. Item 13 (a, b, c)

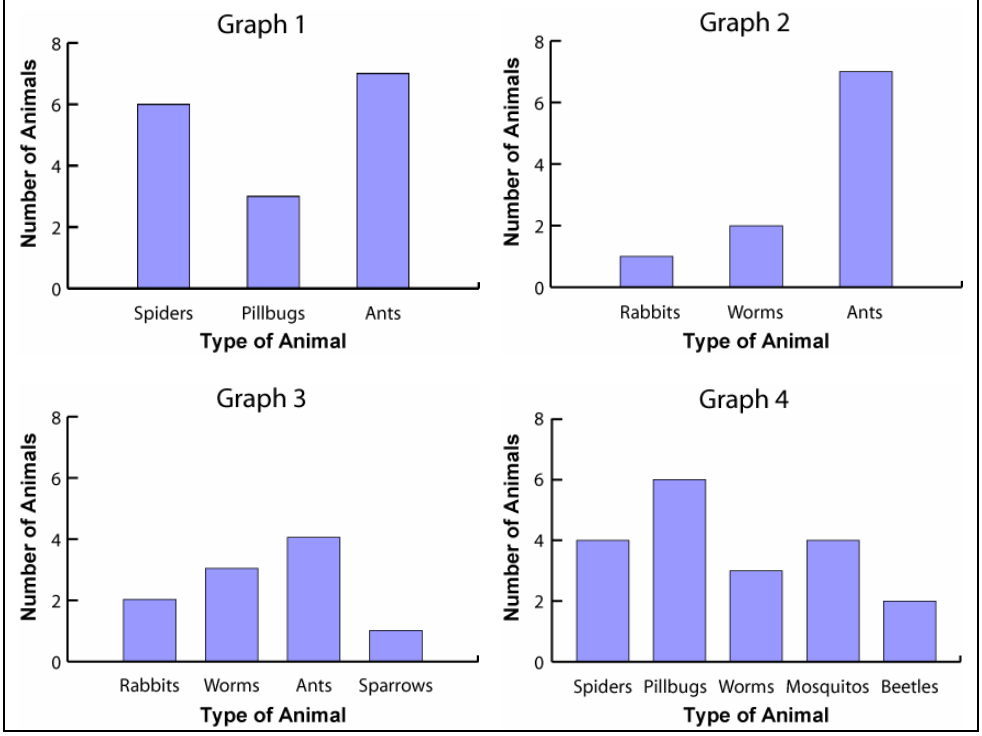
13. Lisa and Juan observed many animals in different parts of their schoolyard. They recorded their observations in the table below:

	Zone A	Zone B	Zone C
Abundance of Animals	30	30	10
Richness of Animals	1	7	3

(a) Which zone of the schoolyard has the greatest biodiversity?
(b) Explain why you chose this zone.

I think that zone _____ has the greatest biodiversity because …

(c) Circle the graph that best represents Zone C



Another outlier is BioKIDS item 11, which is classified as a *step 3 complex* item but on both the pre- and posttests has a below-average difficulty rating. That BioKIDS item 11 was perceived to be easy (with a below-average difficulty even on the pretest) is somewhat surprising. BioKIDS item 11 is found in Table 2 and is the second question dealing with the

scenario of what would happen to the pond system if all of the small fish died one year from a disease. Item 10 asks what would happen to the algae in this situation, and item 11 asks what would happen to the large fish. To answer both of these questions, students must understand the dynamics of a pond ecosystem and the food web interactions involved, and be able to construct a claim and provide evidence to back up their claim without any scaffolding. Item 10 has an above-average difficulty level in both the pretest and the posttest (although not as high as we may have expected) despite seeming similar to item 11 in both content and scaffolding of explanation formation. With such similar cognitive skills involved, it is not easy to reason why these two items have such different difficulties. One possibility is that students are better able to reason "up" a food chain in item 11 (if small fish die, big fish will not have food and will die) rather than "down" a food chain in item 10 (if small fish die, nothing will eat the algae, which will grow more quickly because they are missing a predator). Further investigation of these questions is needed to pinpoint the reason why two items based on the same scenario (in an item bundle), requiring similar content, and having similar format have such different difficulty levels. Having students participate in "think-alouds" would be one way to illuminate where their thought processes differ while solving these questions.

Both of these outliers seem to point to the likelihood that the difficulty of the content knowledge is more indicative of the difficulty of the task than are the format of the question and the amount of scaffolding and inquiry skills provided. In item 13c, students only had to choose a graph; however, the difficulty may have arisen when they were unable to translate terms in the table (richness and abundance) into terms on the graph (types of animals and number of animals). In addition, in item 11, students had to both construct a claim statement and back it up with evidence without scaffolding; however, they found this question easier than other questions of the same format, perhaps because of the content knowledge involved in the question was less difficult. Although our interpretation of these outliers points to content as the key component of difficulty, we cannot statistically make this claim. Because we have based the grouping of our measures on both inquiry skill and content knowledge, we cannot psychometrically tease out whether students found the content or the inquiry process more challenging. Although it would be interesting to determine which aspect of the task gives students the most difficulty, it is unclear whether we want to separate inquiry skill from content knowledge. It might be beneficial to write questions with high levels of content and low levels of inquiry (multiple-choice, fact-based questions) if we are interested solely in students' development of content knowledge. However, although we may be able to write questions requiring high levels of inquiry skills and low levels of content knowledge, we are not sure whether this is something we want to do. Performing inquiry tasks with no content involved does not seem to be very meaningful. Ideally, at higher levels, students can use their content knowledge to help them inquire about scientific issues and come up with explanations based on their inquiry skills. Having students interpret meaningless data or create explanations using unimportant evidence is not the goal of inquiry-based science. Therefore, we want to base our assessment tasks on the types of knowledge that are considered important in our cognitive framework. Although we would like to discover how students improve their content knowledge separately from how they improve on

inquiry, we have to acknowledge that inquiry skills and content knowledge are not independent of each other and therefore perhaps should not be assessed as such.

8.1 Discrimination of Tasks

In addition to allowing us to examine the difficulty of the questions, item maps also point to where items and respondents are aligned on the continuum. When items and students are aligned, the item is a good match to the ability level of the respondent. Ideally, we would like to have items and respondents matched on the continuum so that each respondent would have one or more items that are well suited to measure and distinguish his or her ability level. As is discussed in the results, for the pretests and posttests, we have nonaligning students and items at opposite ends of the continuum (at the lower end for the pretest and the upper end for the posttest). This means that if we want to match the ability levels of our students, we need to develop more easy questions for the pretest and more difficult questions for the posttest. Developing good test questions is not easy; however, our new tools should be able to guide us as to what kinds of questions we need to focus on creating. Even though the mapping of questions did not exactly match our cognitive framework, we can use our matrix to create new questions that are better suited to the ability levels of our students. We seem to have a good range of questions at present; however, in our creation of new tasks, we should focus especially on creating tasks at either end of the difficulty spectrum in order to discriminate accurately between students' ability levels.

8.2 Limitations of This Analysis

For the difficulty analysis, it is possible that a sample of 100 students from the group of more than 2,000 is not a sufficient sample to get accurate data. However, because the 100 students were randomly sampled, they should be representative of the whole group. With more powerful software, running the whole group of students should be easier, and the difficulty parameters and other information should be more reliable.

8.3 Benefits of This Analysis

Despite a few inconsistencies, the data on item difficulty show a pattern whereby students found increasing levels of inquiry and content more difficult. In addition, our tasks appear to be well matched to the ability levels of the students participating in the BioKIDS program. This consistency shows that the cognitive theory underlying our assessment system is well matched with observations of student scores. In the past, we made educated guesses about the difficulty and appropriateness of our assessment tasks for our students; however, with a suite of tasks based on an articulated cognitive theory, this kind of interpretive analysis allows us to determine accurately how well our questions are doing in assessing a range of student knowledge. Especially with students' first foray into science inquiry, it is important to have a continuum of tasks to measure their developing skills. This interpretive analysis shows us what kinds of assessment tasks we need to work on to accurately capture our students' developing inquiry abilities. Using newer PADI tools, such as *templates* and *task specifications*, we could manipulate our assessment items relatively easily; and, with a few changes to some of the items, we will be able to have a more valid and reliable suite of assessment tasks that will allow us to make powerful claims about student learning.

9.0 Conclusion

Too often in assessment development for science inquiry curricula, the entire assessment argument is not fully articulated from the beginning. For science inquiry curriculum developers, the cognitive framework may be known implicitly but never be fully articulated. Without a fully articulated cognitive theory, the tasks that are used or created may not accurately address the knowledge, skills, and abilities that are valued, making it difficult to make strong claims about learning. In addition, the interpretive framework is often very naïve, producing scores that may or may not be reliable and valid. The assessment system created by the PADI group and implemented in the BioKIDS project has combined modern ideas in cognition and measurement to create tools that serve as guides in the creation of science inquiry assessments. The PADI system, however, is not an automatic authoring system and does not take all the work out of creating science inquiry assessment tasks. Rather, PADI provides tools that lead us through a process of articulating the cognitive framework underlying our assessment system, methodically creating our items based on this framework, piloting the items, and then working with the results to revise items as needed. This systematic process allows us to create a strong assessment system, and because one of the main goals of the BioKIDS grant is to longitudinally track students' inquiry skills as they participate in multiple curricular units, having a strong assessment system that is effective in measuring inquiry skills is an essential component of our project.

References

Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A pattern language: Towns, buildings, construction*. New York: Oxford University Press.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17*, 205-226.

Black, P. (2003). The importance of everyday assessment. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 1-11). Arlington, VA: NSTA Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school* (1st ed.). Washington, DC: National Academy Press.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-41.

Driver, R., Guesne, E., & Tiberghisien, A. (Eds.). (1985). *Children's ideas in science*. Philadelphia: Open University Press.

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

Gotwals, A. W., & Songer, N. B. (2004). *A systematic scheme for measuring inquiry skills across curricular units.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Huber, A. E., Songer, N. B., & Lee, S.-Y. (2003, April). *A curricular approach to teaching biodiversity through inquiry in technology-rich environments.* Paper presented at the annual meeting of the National Association of Research in Science Teaching (NARST), Philadelphia.

Krajcik, J., Blumenfeld, P., Marx, R., Bass, K. M., Fredericks, J., & Soloway, E. (1998). Middle school students' initial attempts at inquiry in a project-based science classroom. *The Journal of the Learning Sciences, 7*(3 & 4), 313-350.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*(4), 674-689.

Lee, H.-S. (2003). *Scaffolding elementary students' authentic inquiry through a written science curriculum.* Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Lee, H.-S., & Songer, N. B. (2003). Making authentic science accessible to students. *International Journal of Science Education, 25*(8), 923-948.

Lunetta, V. N. (1998). The school science laboratory: Historical perspectives and contexts for contemporary teaching. In B. J. Fraser & D. Tobin (Eds.), *International handbook of science education* (pp. 249-264). The Netherlands: Kluwer.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Metz, K. E. (2000). Young children's inquiry in biology: Building the knowledge bases to empower independent inquiry. In J. Minstrell & E. H. van Zee (Eds.), *Inquiring into inquiry learning and teaching in science*. Washington, DC: AAAS.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Technical Report 632). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA.

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

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report 9). Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1998, November). *On the role of task model variables in assessment design*. Paper presented at the conference Generating Items for Cognitive Tests: Theory and Practice, Princeton, NJ.

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

Mislevy, R. J., Wilson, M. R., Ercikan, K., & Chudowsky, N. (2002). *Psychometric principles in student assessment* (CSE Technical Report 583). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA.

Mistler-Jackson, M., & Songer, N. B. (2000). Student motivation and Internet technology: Are students empowered to learn science? *Journal of Research in Science Teaching, 37*(5), 459-479.

National Research Council. (1996). *National Science Education Standards*. Washington, DC: Author.

National Research Council. (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning*. Washington, DC: Author.

National Research Council. (2001). *Classroom assessment and the National Science Education Standards*. Washington, DC: Author.

Pellegrino, J. W. (2001). *Rethinking and redesigning education assessment: Preschool through postsecondary*. Denver, CO: Education Commission of the States.

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

Polti, G. P. (1977). *The thirty-six dramatic situations* (L. Ray, Trans.). Boston: The Writers, Inc. (Original work published 1868)

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

Schafer, W. (2002, August). *Describing assessments for teaching and learning.* Paper presented at the conference on Optimizing State and Classroom Tests: Implications of Cognitive Research for Assessment of Higher Order Reasoning in Subject-Matter Domains, University of Maryland, College Park.

Shavelson, R. J., Ruiz-Primo, M. A., Li, M., & Ayala, C. C. (2003). *Evaluating new approaches to assessing learning* (CSE Report 604). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA.

Songer, N. B., Gotwals, A. W., Bao, H., Haertel, G., Hamel, L., Kennedy, C., et al. (in press). *An illustration of PADI design capability in the BioKIDS project* (PADI Technical Report 13). Menlo Park, CA: SRI International.

Songer, N. B., Lee, H.-S., & McDonald, S. (2003). Research towards an expanded understanding of inquiry science beyond one idealized standard. *Science Education, 87*(4), 490-516.

Songer, N. B., & Wenk, A. (2003, April). *Measuring the development of complex reasoning in science.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Van den Berg, E., Katu, N., & Lunetta, V. N. (1994). *The role of "experiments" in conceptual change.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Anaheim, CA.

von Glasersfeld, E. (1998). Cognition, construction of knowledge, and teaching. In M. R. Matthews (Ed.), *Constructivism in science education* (pp. 11-30). The Netherlands: Kluwer.

White, B., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3-118.