The Symbiosis of Cognition, Observation, and Interpretation in an Assessment System for BioKIDS

Amelia Wenk Gotwals and Nancy Butler Songer
The University of Michigan

Abstract

Assessment is a key topic especially 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 However, many current assessments are not good measures of student understanding because they are based on outdated theories of how students learn and only test fact-based knowledge.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 paper examines how an assessment system for the BioKIDS curricular program has been developed based on the three corners of the assessment triangle: cognition, observation, and interpretation. It specifically outlines the theories and beliefs of learning that underpin the system, describes tools created to translate this cognitive framework into tasks that elicit observations of important student knowledge, and uses data to interpret whether the suite of tasks are reliable and predictive of student knowledge. The system provides important information about the BioKIDS assessment system that can inform further iterations of test development and can give credence to the validity and reliability of the test data used to examine student learning.

Introduction

With the shift in pedagogy and learning toward an inquiry-based method of teaching and learning (National Research Council, 1995), the types of knowledge that are

valued has 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 have focused mainly on content knowledge; however, if assessments only examine 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 in both 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 advances both in science teaching/learning and in measurement capabilities. The BioKIDS: Kids Inquiry of Diverse Species project (Nancy Butler Songer, 2000) has teamed with the Principled Assessment Design for Inquiry (PADI) project in order to develop a support structure for the creation, implementation, and scoring of science inquiry assessment tasks. In this paper, I will look at the cognitive theory that is the basis of this assessment system, the methods we have used to translate our cognitive theory into actual assessment tasks, and what the interpretation of the observed student responses tell us about our assessment system as well as about student learning. In particular, I will structure this paper around how the creation and implementation of our systematic assessment program maps onto the three corners of the assessment triangle: cognition, observation, and interpretation – see figure 1 (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) O*bservation* 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 (Mislevy, 2003; Pellegrino et al., 2001). In order to have 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). See table 1 for a summary of how our assessment structures map onto the assessment triangle.

The main research questions that I will be addressing are:

- What is the cognitive framework of the BioKIDS/PADI assessment system?

- How is this cognitive framework translated into tasks that elicit observations of inquiry knowledge?

- What does the interpretation of the observed results tell us about the predictive and systematic nature of our assessment system for both students' inquiry and content knowledge?
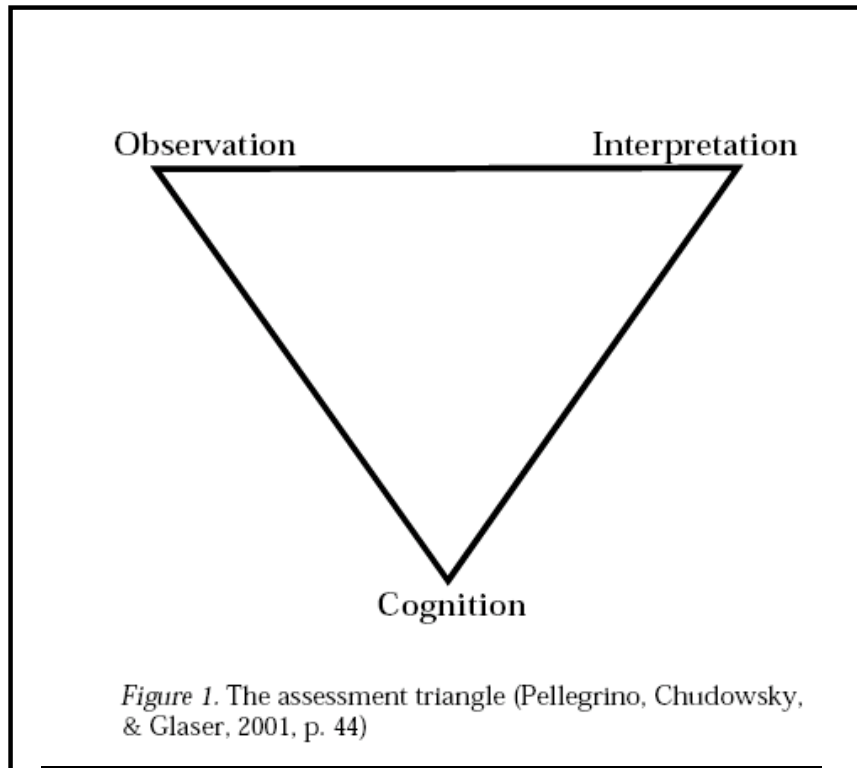
Figure 1. The assessment triangle (Pellegrino, Chudowsky, & Glaser, 2001, p. 44)

Table 1: Map of PADI/BioKIDS Structure to the Assessment Triangle

Assessment Triangle (Pellegrino et al., 2001)	PADI/BioKIDS
Cognition: A set of theory or beliefs about how students represent knowledge and develop competence in a subject	The social constructivist belief system provides the foundational learning theory for scientific inquiry and the complex reasoning and knowledge associated with inquiry. In particular, knowledge, skills and abilities (KSA) that the BioKIDS program emphasizes are focused around three of PADI's design patterns: "formulating scientific explanations from evidence", "interpreting data", and "making hypotheses and predictions".
Observation: A set of beliefs about the kinds of tasks or situations that will prompt students to say, do, or create something that demonstrates important knowledge, skills, and competencies	We have developed Content-Inquiry Matrices representing different complexity levels of both content and inquiry that guide us in the design of assessment tasks. These matrices provide an organizational system to outline the components of tasks that prompt students to demonstrate their understanding of the dimensions of scientific inquiry of interest: formulating scientific explanations from evidence, analyzing data, and building hypotheses and predictions.
Interpretation: A set of assumptions or models for interpreting evidence collected from the observations and all of the methods and tools used to reason from the observations	Coding keys and rubrics provide models of interpreting observations of students work products. In addition, both classical test theory as well as Item Response Theory (IRT) will be used to interpret the observations that out tasks elicit around the dimensions of inquiry reasoning: "formulating scientific explanations from evidence", "interpreting data", and "making hypotheses and predictions".

Cognition

The first corner of the assessment triangle is cognition which outlines the learning theory that drives the assessment as well as the kinds of knowledge, skills and abilities on which this learning theory places value. The cognitive framework of the BioKIDS/PADI assessment system is rooted in two main ideas: Inquiry-based science and constructivist-based assessment theory.

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.

> 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 how scientists study the natural world. (National Research Council, 1995)

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. Part of why inquiry is important is because it can provide students with "real" science experiences, e.g. experiences with many important features of science as practiced by professional scientists (Brown, Collins, & Duguid, 1989) According to the National Science Education Standards (National Research Council, 1995),

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 and conducting 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 paints a very different picture than what can be seen in traditional science classrooms – whether elementary, secondary, or post-secondary. 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 formulae, 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 often emphasizes experiences with fundamental scientific phenomena through direct experience with materials; by consulting books, resources, and experts; and debate among participants (National Research Council, 2000). Inquiry-based learning goals emphasize high expectations including understanding 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 also promote greater student achievement gains in both scientific content and inquiry knowledge (Krajcik et al., 1998; Mistler-Jackson & Songer, 2000; Nancy B. Songer, Lee, & McDonald, 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, p. 31)

BioKIDS: Kids Inquiry of Diverse Species

Research-based curricular development focused on the translation of the science standards into classroom practice, has resulted in several inquiry-focused curricular programs for middle school students. However, 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, 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, & Lunetta 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 scaffold such as prompts, sentence starters or hints about different aspects of inquiry (Lee, 2003).

BioKIDS: Kids' Inquiry of Diverse Species (Songer, 2000) is an IERI-funded project 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 BioKIDS curriculum students participate in is an eight 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. Lee (2002) 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 support in this aspect of inquiry. For many students, this will be their first foray into inquiry-based science learning. As 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 (Nancy Butler Songer & Wenk, 2003). We wanted to be able to see students' progression through both a single curricular unit as well as across curricular units and determine their level of reasoning ability at each stage.

In order to gain the tools necessary to follow students' learning trajectories as they participate in the BioKIDS curriculum, the BioKIDS project has joined the PADI team to create inquiry assessments. The PADI project's main focus is the development of a conceptual framework and support structure for the systematic development and

implementation of assessment tasks associated with measuring scientific inquiry. PADI combines developments in cognitive psychology, research on scientific inquiry and advances in measurement theory and technology to formulate a structure for developing systematic inquiry assessment tools. Experts in each of these fields contribute to the process of developing this system, and in doing so, help to provide common terminology and design guidelines that make the design of an assessment explicit and "link the elements of the design to the processes that must be carried out in an operational assessment" (Mislevy, 2003).

Constructivist-Based Assessment

All assessments are based in a conception/philosophy of how people learn; of what tasks are most likely to elicit observations of knowledge and skills from students; and are premised on certain assumptions about how best to interpret evidence to make inferences (Mislevy, 2003). Many assessment that are being used today are created 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 which 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).

> 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…
(Pellegrino, 2001)

The current theory of learning that many subscribe to, and that inquiry science 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, 2000; Driver, Guesne, & Tiberghisien, 1985; von Glaserfeld, 1998). The kinds of assessments that are based on constructivism will likely be considerably different than those based on behaviorism. Assessments in line with constructivist theories of learning should 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 classrooms (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, including selection/placement, accountability for schools and teachers, as well as measurement of student learning (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, 2003). One of the key steps in assessment is the actual design of tasks. In the past, task design was seen more as an art than a science, however, a more principled approach to assessment task

design is needed 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).

Observation

The second corner of Pellegrino's assessment triangle is observation. The observation corner refers to the "kinds of tasks or situations that will prompt students to say, do, or create something that demonstrates important knowledge and skills" (p. 47). The tasks used to systematically assess inquiry skills cannot be randomly created or chosen from a set of tasks, rather they must be tightly linked to the cognition corner of the triangle to ensure that the knowledge and skills that are valued are the knowledge and skills that are being tapped. Creating assessment items that accurately measure certain skills or abilities is not an easy task. Having tools can help to make the transition from cognitive framework to actual task. The creation of these tools is an essential component of the PADI design system.

Translating the Cognitive Framework into Tools to Create Tasks

Design Patterns for Inquiry

While science standards documents (National Research Council, 1995, 2000) outline aspects of inquiry that are important for students to learn, they do not provide a

cohesive guiding structure to assess these skills. The PADI team has developed

structures that provide guidance in translating standards and curricular learning goals into

assessment tasks that reliably measure scientific inquiry skills. The large umbrella

structures under which all of PADI assessment task design falls are design patterns

(Mislevy et al., 2003).

Design patterns have been used in other disciplines for many years. The design

patterns that have been created and utilized 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 *The Thirty-Six Dramatic Situations*. In 1868, Polti claimed that all

literary works are based on and can be categorized into thirty-six dramatic situations such

as "Falling prey to cruelty or misfortune" and "Self sacrifice for kindred" (Mislevy et al.,

2003; Polti, 1868). The second situation, "Self sacrifice for kindred," is found in plays

and novels such as Shakespeare's *Measure for Measure*, Rostand's *Cyrano de Bergerac*,

Dickens' *Great Expectations*, and Edith Wharton's *Ethan Frome* (Mislevy et al., 2003).

Clearly, these works are not identical, they tell very different stories. However, they all

possess a clearly delineated combination of elements: the hero or protagonist, a family

member in need, and some sacrifice made by the hero in order to alleviate this need.

Much can be varied within this design, such as what is sacrificed, the reasons for the

sacrifice, and the relationships between key characters, but these elements stay the same

(Mislevy et al., 2003).

Polti explicated these situations or themes both to show similarity between

dramatic stories as well as 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 be used as tools to help authors generate new stories (Mislevy et al., 2003).

The concept of design patterns is also 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 as both a tool for analyzing or classifying preexisting artifacts (such as buildings or programs) as well as providing 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 her 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 (Risconscente, Mislevy, & Hamel, in press).

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. In order 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 et al., 2003). 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).

In the BioKIDS project we focus on three main design patterns: "formulating scientific explanations using evidence," "interpreting data," and "making hypotheses and predictions." To illustrate how design patterns can be used as a tool, I will focus on the design pattern, "formulating scientific explanations using evidence." "Formulating scientific explanations using evidence" is one of the key aspects of inquiry that the BioKIDS curriculum fosters using direct scaffolding. Table 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 five essential features of classroom inquiry, making it an "essential essential" (National Research Council, 2000). Explaining the importance of the inquiry skill appears in the rationale section of the design pattern table. In line with the assessment triangle discussed above, the design pattern table provides space to list the focal knowledge, skills and abilities (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 explanations using 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 assessment tasks.

The design pattern table also provides room to articulate aspects of a task (characteristic features) that would elicit the observations needed as evidence of the KSAs, as well as work products that could employ these features. For example, since we define an explanation consisting of a claim and use of evidence to back up the claim (Kuhn, 1989; Toulmin, 1958), observations we may 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 kind 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 be a multiple-choice question, while a harder question could involve a student building an explanation without guidance. While these tasks have different formats, they share certain characteristic features such as including both a claim and evidence.

Table 2: Design Pattern for "Formulating Scientific Explanations from Evidence"

Attribute	Value(s)	Comments
Name	Formulating scientific explanation from evidence	
Summary	In this design pattern, a student develops a scientific explanation using the given evidence. The student must make a relevant claim and then justify the claim using the given evidence.	A scientific explanation consists of stating a claim and using the given data appropriately to support this claim. A scientific explanation is different from other explanations because it requires using relevant evidence to justify it.
Rationale	Two key aspects of scientific inquiry are the ability to understand scientific phenomena and the ability to be able to propose explanations using evidence. This design pattern addresses both of these.	The National Research Council lays out five essential features of classroom inquiry. Four of the five aspects involve students using evidence to create and justify explanations.
Focal KSAs	The ability to develop scientific explanations using evidence.	
Additional KSAs	• Conducting appropriate inquiry practices for the scientific question at hand. • Weighing and sorting data/evidence. • Formulating a logical claim according to the given data/evidence.	
Potential Observations	The claim reflects an understanding of the data given and a certain amount of scientific knowledge	The amount of scientific knowledge involved can vary depending on the level of the assessment item
	There should be logical consistency between the evidence and the claim	Does the evidence actually back up the claim
	The data that is used to support the claim is relevant and the more pieces of relevant data used.	
Characteristic features	Item provides space for claim and data/evidence	
Variable features	Level of prompting	Less prompting makes the item more difficult for the student and thus gives better evidence about whether student is able to provide scientific explanations using data on their own. More prompting makes the item easier and thus gives evidence about whether a student is able to provide an explanation using data when given the appropriate format in which to do so.
	Difficulty of the problem context/content	The level of the question can be varied by the amount of content the student needs to bring to the question as well as the amount of interpretation of the evidence is necessary.

	Amount of evidence provided	The amount of evidence provided can make the question easier or harder. If more irrelevant information is provided, students will have to be better at sorting to find the appropriate evidence to use. However, if more relevant information is provided, finding evidence to support a claim will be easier.

Content-Inquiry Matrix

Although the tasks formulated using 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 et al., 2003). Tasks focused on 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 (Nancy B. 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 section 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 the difficulty of science inquiry assessment tasks as having two dimensions: the difficulty of the science content and the difficulty of the science inquiry. To address both of these aspects of task difficulty, we created a matrix that lays out three possible levels for each dimension (see Table 3). First we classified science content knowledge into: **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 be able to link different concepts together. Secondly, we separated inquiry into three levels: step1, step 2, and step 3. While the content aspect of the matrix can remain the same or very similar for all design patterns, the steps of

inquiry will vary due to the inherently different nature of the aspects of inquiry being targeted. For the "formulating explanations from evidence" design pattern, we borrowed from our curricular units and created degrees of inquiry tasks based on the amount of support or scaffolding the task provides for explanation formation. **Step 1** tasks provide evidence and a claim, and students simply need to match the appropriate evidence to the claim (or vise versa). While 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 choice. This involves more inquiry ability than the step 1 task of matching, but there is still support for students guiding 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 in 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 have also created similar matrices for the other two design patterns that we focus on: interpreting data and making hypotheses and predictions.

In the past, science performance assessments have been classified based on the amount of content involved in and the freedom given to students in conducting scientific investigations (Baxter & Glaser, 1998). In particular, Baxter and Glaser identify four quadrants that performance assessment tasks can fall into based on the amount of content involved (content-rich or content-lean) and the amount of freedom students are given with regards to process or inquiry skills (constrained or open). Our matrix looks at a

similar dimension of content and a different dimension of inquiry. In addition to just the

amount 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/or the interrelationships of these processes. Our matrix also examines the

amount of inquiry required to solve the task. The main difference between our matrix

and Baxter and Glaser's quadrants is that our matrix is specific to a single inquiry skill or

design pattern (like formulating explanations using 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 they do

not separate out different aspects of inquiry skills (design patterns). Instead, they group

all skills involved in the investigation together and measure inquiry as a whole. Both of

these tools are useful in characterizing inquiry assessment tasks. Our matrix is more

likely to be used create tasks that measure specific inquiry abilities, whereas, Baxter and

Glaser's quadrants are more useful for designing and classifying performance

assessments where students conduct full scientific investigations.

Table 3: Levels of Content and Inquiry Knowledge Needed for Assessment Items Related to the Design Pattern: "Formulating scientific explanation from evidence"

	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 choose the 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, in order 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 claim using relevant evidence (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 requires either interpretation or content knowledge	Students must construct a claim and explanation that requires the students to interpret evidence and apply additional content knowledge.

Using the Tools to Create Tasks

Task Design in BioKIDS

Using the structure provided by the design patterns and the content-inquiry matrices, we used both a reverse and forward design process in order 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 design patterns including "interpreting data," "re-expressing data," and "making hypotheses and predictions." Although some previously written items did fall into our matrix, 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.
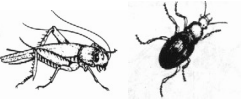
In mapping old tasks and creating new tasks, we used the content-inquiry matrix to make sure that we were examining all levels of content and inquiry knowledge. Most of the tasks fell into one of three categories: Step 1 simple; Step 2 moderate; and Step 3, complex. We found that developing Step 1 simple, moderate, or complex tasks was relatively easy; these tasks were generally multiple-choice questions with varying degrees content difficulty. In contrast, we found it difficult to authentically address high levels of inquiry (both steps 2 and 3) without involving content. 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 development from inquiry skills development. Thus, despite confounding inquiry skills and content understanding with 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**).
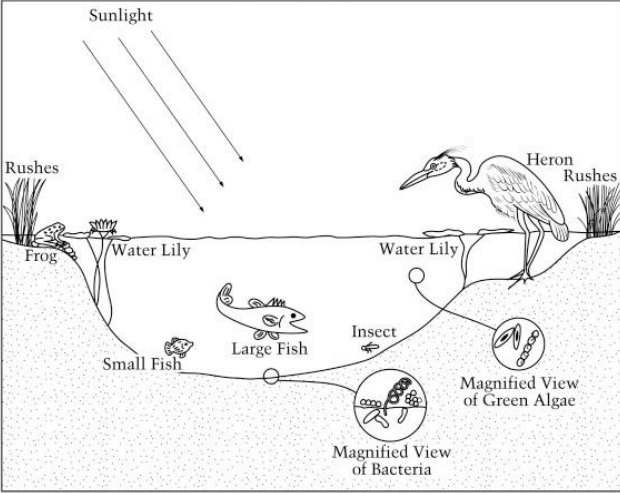
Table 4 provides examples of three tasks from the "formulating scientific explanations" design pattern that fall in the cells along the diagonal in the matrix. As is

clear from both the matrix and the examples, each level up (simple to moderate to complex) requires more (quantity) and more difficult content knowledge. Our first question has a table that provides all of the content information that a student requires in order to complete the task successfully; all the students need to do is to choose the relevant evidence from the table and match it to the provided claim. Our second question provides students with pictures of invertebrates that they must group together based on certain characteristics. Students are provided with the pictures, so they are not required to know all of the physical characteristics that separate insects from arachnids, however, in order to answer the question correctly, they do need to know what physical characteristics are important when classifying animals. Students are provided with a pre-formed claim statement where they just have to choose the appropriate claim, and then they are prompted to give evidence. Finally, our last question 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 using evidence, the tasks are clearly of different levels. Having these different levels is important if we want to measure students' developing inquiry abilities. If we only had step 1, questions (only multiple choice questions), we would not be able to see if students could progress past the stage of matching claims and evidence. On the other end, if we only had step 3 questions, we would not be able to determine if students hold more tenuous explanations building skills, 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 4: BioKIDS Questions Mapped to the Level of the "Formulating Scientific Explanations Using Evidence" Design Pattern

Question	Step and 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

	Step 3, Complex
POND ECOSYSTEM 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. *(From NAEP assessment)*	

<u>Interpretation</u>

Despite having a solid cognitive theory that has guided task design, no assessment can actually "get into" a student's head and measure exactly what they know or can do. Therefore, every assessment must be designed with an interpretation model in mind. This model includes "all the methods and tools used to reason from fallible observations" (Pellegrino et al., 2001). In order to determine students' knowledge, skills and abilities, we administered the BioKIDS test to a group of students and interpreted the results.

Data Collection

In Fall 2003, over 2,000 sixth grade students from sixteen 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 pre and post test made up of sixteen questions, drawing from the diagonal cells of the matrices from each of the

three focal design patterns (formulating explanations using evidence, interpreting data, and making hypotheses and predictions). In order to determine some properties of the assessment tasks, I use the data gathered from both students' pre and posttests.

Methods

Scoring:

While all of the tasks for the BioKIDS assessments were created using the same design pattern with specified levels of content and inquiry, we cannot equate any of the measures, psychometrically, as either parallel or even as measuring the same construct. This is where the scoring and interpretation of scores comes into play. Based on our cognitive model and previous student answers to questions, we developed a coding rubric to score all of 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, we coded students' claim statement separately from their use of evidence to support their claim. Generally, the claim statement was scored 0 if it were incorrect or blank and 1 if it was correct. The evidence portion of the code 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 given (up to two pieces of evidence). Because of the large number of tests, five people (who first established >90% inter-rater reliability) coded all of the tests. For the rest of this paper, I will use these scores to examine some of the basic psychometric properties of the biodiversity assessment.

Difficulty, discrimination and fittedness of the tasks

It is important to compare the mapped difficulty level of a question (Step 1, 2, or 3) with the empirical difficulty level to determine if the cognitive scheme guiding our task creation using design patterns and content-inquiry matrices maps onto what students experienced when they completed the tasks. In order to determine the predictability and accuracy of our cognitive model, I used the student version of the Rasch modeling software Winsteps, which is called Ministeps. I used a Rasch (one parameter) model, and although this model does not take into account factors other than difficulty involved in how students respond to questions, it allows for a good estimate of difficulty level. Because Ministeps has a reduced capacity load, I used SPSS to randomly choose one hundred students who had completed both the pre and the posttest from our database. Choosing students randomly allows me to generalize the results 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, I had to run a 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, I calculated the difficulty parameter and created item maps for both the pre and 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 processes 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.

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, both assessment tasks as well as 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 looking at the position where they fall on the continuum, one can determine the difficulty of the item (if they are close to zero, they have an average difficulty, above zero they are more difficult and below zero, less difficult). On the left hand side of the divider are X's, 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, I focus on difficulty level, so once I created the item maps, I color-coded them by matrix position regardless of design pattern (step 1 simple…).

Empirical Results

Difficulty, discrimination and fittedness of the tasks

There are two difficulty tables and item maps, one for the pretest and one for the posttest. Tables 6 and 7 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 whereas the range for the posttest is smaller with the least difficult items having a difficulty of -2.5 and the most difficult items only having a difficulty of 2.11.

Table 6: Pretest Items Difficulty Listed from Most Difficulty to Least Difficult
(bolded items are outliers)

Item	Estimated Difficulty	Step and 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 6	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

Table 7: Posttest Items Difficulty from Most Difficulty to Least Difficult (bolded items are outliers)

Item	Estimated Difficulty	Step and 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, Complex
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

The item maps (figures 2 and 3 for pre and post-test respectively) give the similar information as is in the tables, only in a graphical structure where both the items and respondents are on the same continuum (θ). Items in red fall in the step 1, simple cell of the matrix, items in green in the step 2 moderate cell of the matrix, and items in blue 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 -4.5) than we have questions for (the least difficult question is placed at -3.3 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 2: Item Map Pretest By Complexity Level ($N_{Students}=100$; $N_{Items}=26$)

```
       KIDS MAP OF Items
         <more>|<rare>
  3      X   +
              |
          XX  |
            T|  BioKIDS14a
          X   |
          X   |   BioKIDS13c   BioKIDS13a
              |
  2      XX   +   BioKIDS16a   BioKIDS6d
          X   |   BioKIDS9
             |S  BioKIDS14b
          X   |
         XX   |   BioKIDS15
          X   |
         XX  S|
  1     XXXX  +
         XXX  |   BioKIDS10
        XXXXX |   BioKIDS8
         XXX  |   BioKIDS13b
          XX  |   BioKIDS14c   BioKIDS6e
      XXXXXXX |
         XXX  |
  0     XXXXX +M  BioKIDS5a
         XXXX |   BioKIDS4a
       XXXXXX M|
        XXXXX |
         XXXX |
           X  |   BioKIDS11     BioKIDS2
         XXXX |   BioKIDS6b
 -1    XXXXXX +
           X  |
          XX  |
           X  |   BioKIDS12     BioKIDS4b
        XXXXX |
            S|S  BioKIDS1
         XXXX |   BioKIDS6c
 -2       X   +
          XX  |   BioKIDS7
           X  |   BioKIDS5b     BioKIDS6a
              |
              |
              |
 -3      XXX  +
             T|
              |   BioKIDS3
             |T
              |
              |
 -4           +
              |
         XXX  |
              |
 -5           +
        <less>|<frequ>
```

Step 1 Simple
Step 2 Moderate
Step 3 Complex
X – Students

Figure 3: Item Map Posttest By Complexity Level (N$_{Students}$=100; N$_{Items}$=26)

```
    KIDS MAP OF Items
      <more>|<rare>
4      X   +
           |
      XX   |
           |
       X   |
           |
           |
          T|
3      X   +
      XX   |
       X   |
           |T
      XX   |
     XXX   |
           |
       X   |   BioKIDS14a
2     XX  S+   BioKIDS16a
      XX   |
           |   BioKIDS6d
     XXX   |
    XXXXX  |
   XXXXXX  |
    XXXXX  |S
     XXX   |   BioKIDS13a   BioKIDS14b   BioKIDS15    BioKIDS9
1   XXXXX  +   BioKIDS13c
     XXX   |
    XXXX  M|   BioKIDS14c
    XXXX   |
       X   |   BioKIDS10     BioKIDS13b   BioKIDS8
    XXXXX  |   BioKIDS2      BioKIDS5a
    XXXX   |
      XX   |
0    XXXX  +M  BioKIDS4a
    XXXXX  |   BioKIDS6b     BioKIDS6e
    XXXX   |
      XX  S|
     XXX   |
    XXXX   |
       X   |
      XX   |
-1     X   +
       X   |   BioKIDS11
      XX   |S  BioKIDS5b
           |   BioKIDS12     BioKIDS1      BioKIDS7
       X   |   BioKIDS4b
         T|
           |   BioKIDS6a
           |
-2         +   BioKIDS6c
       X   |
           |
           |   BioKIDS3
          |T
           |
           |
-3         +
      <less>|<frequ>
```

Step 1 Simple
Step 2 Moderate
Step 3 Complex
X – Students

Discussion

Difficulty and fittedness of the tasks

The third research question that I posed was, "What does the interpretation of the observed results tell us about the predictive and systematic nature our assessment system for both inquiry 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 having 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 a few "outliers" within the each of the categories (step 1 simple, step 2 moderate, and step 3 complex items). I now focus on two of 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, and yet is the second most difficult item on the pretest and has over 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 this question in table 8). In the BioKIDS program we focus on three main design patterns: formulating scientific explanations using evidence; interpreting data; and making hypotheses and predictions. This question fits into a separate design pattern called "re-expressing data." This question is part of an item bundle where the other items are focused on interpreting data and formulating explanations. This section of the question has students recognize that the

same data can be expressed in many forms – in particular for this case, a table and a graph. Though this skill is addressed in the 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 where students have to construct their own responses. However, according to the interpretation of results, students did not perceive this item to be easy.

There are several factors that could make this item more difficult than we originally thought. On careful examination of this question, there are several cognitive steps that students must go through in order to successfully answer this question. First of all, 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. Secondly, 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 animal in the graph. In addition, the graph introduces names of animals that are not found anywhere in table. Students need to recognize that they do not need to look for specific animal names; rather they are looking for total number of animals 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 re-expressing data is simply a difficult skill for students of this age, and more scaffolding may be needed in order to make it 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.
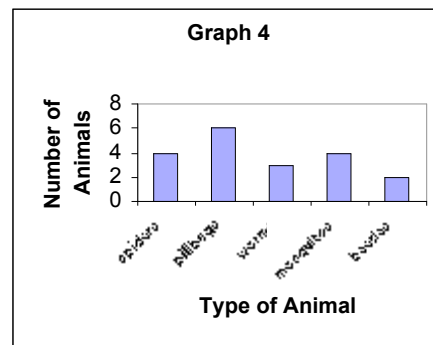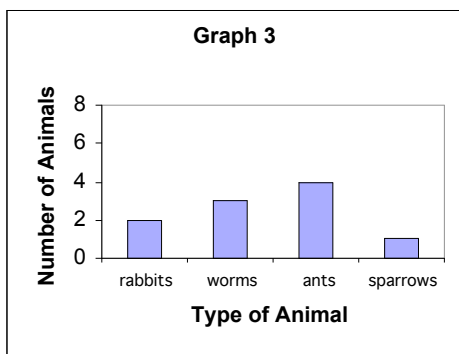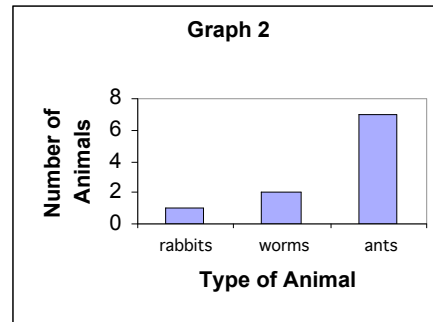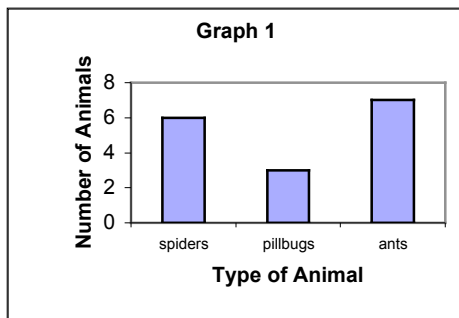
Table 8: Item 13(a, b, c)

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 &B) Which zone of the schoolyard has the greatest biodiversity? Explain why you chose this zone.

I think that zone _____ has the greatest biodiversity because …

(C) Circle the graph that best represents Zone C

Graph 1 — Number of Animals vs Type of Animal (spiders, pillbugs, ants)

Graph 2 — Number of Animals vs Type of Animal (rabbits, worms, ants)

Graph 3 — Number of Animals vs Type of Animal (rabbits, worms, ants, sparrows)

Graph 4 — Number of Animals vs Type of Animal

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 3 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. In order to answer both of these questions, students have to understand the dynamics of a pond ecosystem, the food web interactions involved, as well as have to 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 pre and the posttest (although not as high as we may have expected) despite seeming very 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 (that 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 and it will grow more quickly because it is missing a predator). Further investigation of these questions is needed in order 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 process differ while solving these questions.

Both of these "outliers" seem to point to the fact that perhaps the difficulty of the content is more telling of the difficulty of the task than the format of the question and the scaffolding of inquiry skills. 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 (number of animals…). In addition, in item 11, students had to construct both 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 the content involved in the question was less difficult. While my interpretation of these outliers point to content being the key component of difficulty, we cannot statistically make this claim. Because we have grouped our measures based on both inquiry and content, we cannot psychometrically tease out whether students found the content or the inquiry more challenging. While 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, while we may be able to write questions requiring high levels of inquiry skills and low levels of content, we are not sure if this is something we want to do. Performing inquiry tasks with no content involved does not seem to hold much meaning. 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. 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. While we would like to discover how students improve in content separate from how they improve on inquiry, we have to acknowledge that inquiry and content matter are not independent of each other and therefore should not be assessed as such.

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 their ability level. As is discussed in the results, for the pre and posttests we have non-aligning students and items at opposite ends of the continuum (on 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 kind 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 the present; however, in our creation of new tasks, we should especially focus on creating

tasks at either end of the difficulty spectrum in order to accurately discriminate between students' ability levels.

Limitations of this analysis:

        For the difficulty analysis, it is possible that a sample of 100 students from the group of over 2000 is not a sufficient sample to get accurate data. However, since the 100 students is a random sample, it 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.

Benefits of this analysis:

        Despite a few inconsistencies in our expected results, the main pattern in the data in terms of difficulty shows that students found increasing levels of inquiry and content more difficult. In addition, our tasks appear to be well matched to the ability level 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 interpretation analysis allows us to accurately determine how well our questions are doing in assessing a range of student knowledge. Especially with students' first foray into inquiry science, it is important to have a continuum of tasks to measure their developing skills. This interpretative analysis shows us what kinds of assessment tasks we need to work on to accurately capture our

students' developing inquiry abilities. With a few changes to some of the items, we will be able to have a valid and reliable suite of assessment tasks that will allow us to make powerful claims about student learning.

Conclusion

Too often in assessment development, the three corners of the assessment triangle are not strongly tied together. For inquiry-science curriculum developers, the cognitive framework may be known implicitly, but may 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 of learning. In addition, the interpretive framework is often very naïve, producing scores which may or may not be reliable and valid. The assessment system that we have created in the BioKIDS/PADI group has combined modern ideas in cognition and measurement in order to create tools that serve as guides in the creation of science inquiry assessments. As 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 a key component of our project. With the knowledge gained from the development of our assessment system and the analysis of the results, we have a test that is valid, reliable, and matched to the ability levels of our students.

Works Cited:

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* (1 ed.). Washington D.C.: 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.

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 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.* 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, Publishers.

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. v. Zee (Eds.), *Inquiring into Inquiry Learning and Teaching in Science*. Washington D.C.: AAAS.

Mislevy, R. J. (2003). *A Brief Introduction to Evidence-Centered Design* (Technical). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Haertel, G., Hamel, L., et al. (2003). *Design Patterns for Assessing Science Inquiry* (Technical Report): SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1998, November 5, 1998). *On the Role of Task Model Variables in Assessment Design.* Paper presented at the Generating Items for Cognitive Tests: Theory and Practice, Princeton, NJ.

Mislevy, R. J., Wilson, M. R., Ercikan, K., & Chudowsky, N. (2002). *Psychometric Principles in Student Assessment*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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. (1995). *National Science Education Standards*. Washington, DC: National Research Council.

National Research Council. (2000). *Inquiry and the National Science Education Standards: A Guide for Teaching and Learning*. Washington, D.C.: National Research Council.

National Research Council. (2001). *Classroom Assessment and the National Science education standards*. Washington D.C.: National Research Council.

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 D. C.: National Academy Press.

Polti, G. P. (1868). The Thirty-Six Dramatic Situations. In Ray L. (translated 1977) (Ed.). Boston: The Writers, Inc.

Risconscente, M., Mislevy, R. J., & Hamel, L. (in press). *An Introduction to PADI Task Templates*. Palo Alto, CA: SRI International.

Schafer, W. (2002, August 15-16, 2002). *Describing Assessment for Teaching and Learning.* Paper presented at the Paper presented at 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*. Los Angeles, CA: Center for the Study of Evaluation (CSE).

Songer, N. B. (2000). *BioKIDS: Kids' Inquiry of Diverse Species*: Proposal funded by the Interagency Educational Research Initiative (IERI) $5.5 million.

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 25, 2003). *Measuring the Development of Complex Reasoning in Science.* Paper presented at the AERA (American Education Research Association), Chicago, IL.

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

von Glaserfeld, 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.