What Constitutes Evidence of Complex Reasoning in Science?

Nancy Butler Songer & Amelia Wenk Gotwals
School of Education, The University of Michigan
Tel: (734) 647-7369, Fax (734) 763-1368
Email: songer@umich.edu

Abstract: The U.S. priority on testing and accountability challenges school administrators to adopt, adapt or develop assessment systems that provide concrete evidence in particular subject matter areas such as science. Curricular programs emphasizing constructivist learning approaches and complex reasoning such as scientific inquiry are often not well matched to the assessment instruments used to evaluate them, such as standardized or off-the-shelf tests. As a result, educational researchers are sometimes torn between supporting curricular activities that promote complex reasoning and supporting high-stakes tests designed to emphasize facts and conceptual knowledge development over higher-order reasoning. This paper presents results from the first year of a research program to develop and evaluate curricular/assessment systems emphasizing complex reasoning in science. Results provide information for new models of assessment systems that complement high-stakes tests as they address the question, what constitutes evidence of complex reasoning in science?

Introduction

Perhaps never before has the issue of measurement of student learning in science and other content areas been so complex and important. While international tests demonstrate that American middle school students' achievement in science declines relative to their peers internationally (e.g. (Linn, Lewis, Tshuida, and Songer. 2000), education reform laws such as the No Child Left Behind Act support higher levels of accountability and larger consequences for poor performance. Concurrently, national organizations such as the American Association of the Advancement of Science (AAAS) advocate standards-based curricular programs to foster the development of complex reasoning in science, including both the ability to explain individual scientific concepts and the relationships and connections between concepts. To accompany these standards-based curricular programs, the National Research Council (National Research Council 2001) recommends robust assessment instruments that provide compelling evidence of students' complex reasoning and relationships among scientific ideas.

"Assessments that resonate with a standards-based reform agenda reflect the complexity of science as a discipline of interconnected ideas and as a way of thinking about the world." (National Research Council 2001; p. 12)

Despite the demand for assessment instruments that measure complex reasoning such as science inquiry, few instruments exist that provide a systematic approach to the evaluation of complex reasoning in science (Mislevy et al, 2002). Many of the current high-stakes national and international science tests emphasize definitions of science concepts and/or fact-based knowledge over items measuring complex reasoning in science, no doubt because of the challenge of developing reliable instruments to systematically evaluate scientific reasoning. As high-stakes tests attempt to match the learning goals of the standards-based reform programs but often fall short, schools are challenged to provide tangible evidence of success on the complex reasoning associated with the reforms through either ill-suited standardized tests or off-the-shelf assessments (Mislevy, in press). Systematic assessment programs are needed that provide valid, tangible evidence of complex reasoning well matched to the learning goals emphasized in the standards-based reform curricula (Pellegrino, Chudowsky, and Glaser, 2001).

Failing schools, such as those in many urban school districts, are particularly pressured to perform well on high-stakes tests. A systematic approach to inquiry assessment might be especially valuable for urban schools to provide tangible evidence of student performance associated with curricular reforms (Songer, Lee, and McDonald. 2002). This paper describes early results obtained through the development of one assessment system designed to evaluate urban students' complex reasoning in science associated with a multi-year, standards-based curricular program implemented in one inner city district.

The Development of Complex Reasoning in Science

Research on children's learning recognizes that the development of deep conceptual understandings in science requires the structuring of experiences, including catalysts to encourage curiosity or persistence and mediation such as scaffolds to guide attention to salient features within complicated reasoning situations (Lee and Songer, 2003; Bransford, Brown and Cocking, 2000; Vygotsky, 1978). The development of complex inquiry thinking in science requires both the development of underlying science concepts as well as the development of reasoning skills in that context, such as building explanations from evidence or data (National Research Council, 2000). In addition, research demonstrates that the development of complex thinking takes time (Bransford, Brown and Cocking, 2000) and is therefore not well suited to short-term curricular interventions. Ideally, children's inquiry knowledge development occurs systematically over multiple coordinated units and years.

While programs fostering the longitudinal development of complex reasoning associated with a particular content area exist in mathematics and other areas, it is interesting that few curricular programs in science are designed to systematically promote students' inquiry knowledge development over multiple units or years. An idealized curricular program would consist of a sequence of curricular units that are matched to a coordinated set of inquiry-focused learning goals, and that systematically build reasoning skills throughout the coordinated units. The learning goals in this idealized curricular program would emphasize both the development of deep conceptual understandings in the content areas as well as the development of complex reasoning skills such as formulating explanations from evidence or analyzing and interpreting scientific data.

Over the past eight years, The BioKIDS: Kids' Inquiry of Diverse Species project [www.biokids.umich.edu] has developed and evaluated two eight-week curricular units to foster complex inquiry thinking for sixth graders around topics in biodiversity and weather. The National Research Council (National Research Council, 2000) lays out five essential features of classroom inquiry and the variety of ways that they can be seen in practice. Four of the five aspects involve students using evidence to create and justify explanations, making explanations a large component of classroom inquiry. Thus, one of the main aspects of scientific inquiry that the BioKIDS curricular programs promote is students' development of scientific explanations using evidence.

In these units, particular inquiry thinking skills, such as the formulation of scientific explanations from evidence, are fostered through a carefully scaffolded activity sequence (Songer, 2003; Huber, Songer and Lee, 2003). One characteristic of the curricular sequence is the collection and analysis of data using emerging technological tools (Parr, Jones and Songer, in press) accompanied by the repeated presence of guided-learning prompts to support particular types of complex thinking associated with the scientific data. For example in the biodiversity unit, sixth grade students use PDAs loaded with animal tracking software (CyberTracker; see (Parr, Jones and Songer, in press) to collect animal species data on a particular zone of their schoolyard. Once data are collected, students use their animal data to formulate scientific claims and evidence-based explanations to address the question, "Which zone in the schoolyard has the greatest biodiversity?" Scientists might evaluate which zone is most diverse using Simpson's index, $D = 1-E (n/N)^2$, a formula that represents species evenness taking into account both the total number of animals (abundance) and the number of different species (richness). In BioKIDS, students develop a qualitative response in the form of a claim and evidence that, like Simpson's index, also takes into account species abundance and richness. Twelve times throughout the biodiversity unit, students use guided prompts and their own scientific data to formulate scientific explanations based on evidence. This example illustrates some of the dimensions of the curricular program that promote inquiry reasoning focusing on "formulating explanations from evidence" throughout the biodiversity unit. In the longitudinal study that will build from this study, guided prompts will encourage cohorts of students to formulate explanations from evidence in several other coordinated curricular units in the 6th, 7th and 8th grades.

A Coordinated Assessment System for Measuring Scientific Inquiry

Having assessment tools that are able to follow students' learning trajectories as they participate in each of the inquiry units is essential to provide empirical evidence of students' abilities to perform complex reasoning skills. In order to develop these assessment tools, BioKIDS participates in the IERI Principled Assessment Design for Inquiry (PADI) project [http://padi.sri.com/]. The main focus of PADI is the development of a conceptual framework and a support structure for the systematic development and implementation of assessment tasks for 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. Standards documents such as the National Science Education Standards (National Research Council, 1996) and Inquiry and the National Science Education Standards (National Research Council, 2000) outline aspects of inquiry that are important for students to learn, however, they do not provide a guiding structure to help assess these skills. PADI provides structures that provide translations between standards and curricular learning objectives of a particular curricular reform program, corresponding assessment tasks, and the measurement models that provide information on the evaluation of scientific inquiry skills.

The PADI project focuses on the systematic development of several pieces of a support structure that bridge between learning goals, assessment tasks, and measurement models focusing on scientific inquiry. One of the foundational units of the PADI system is a support feature called "design patterns". Design patterns consist of a narrative matrix focused around concrete inquiry learning goals such as "formulating scientific explanations from evidence". This narrative matrix outlines "the chain of reasoning, from evidence to inference" (Mislevy et al, 2002) through information in three areas: (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 2002). The design pattern for "formulating scientific explanations from evidence" links assessment goals, science standards, and curricular learning objectives with appropriate assessment task models and formats designed to evaluate students' explanations. Through the development of the structural piece called a design pattern, the PADI team has begin the systematic work necessary to create a series of assessment tasks that can accurately measure some of the complex reasoning skills demonstrated in inquiry-based science classrooms. Table 1 presents selected items from the design pattern, "formulating scientific explanations from evidence."

Table 1: Design Pattern Matrix, Selected Items, for "Formulating Scientific Explanations from Evidence"

Attribute	Value(s)
Name	Formulating scientific explanation from evidence
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. View the situation from a scientific perspective
Potential Observations	The claim reflects an understanding of the data and a certain amount of scientific knowledge
	There should be logical consistency between the evidence and the claim
	The data that is used to support the claim is relevant and the more pieces of relevant data used
Potential Work Products	Multiple Choice - matching claim and evidence Spoken explanation when in a situation involving scientific concepts Written response - creation of claim and use of appropriate evidence to justify claim
Characteristic features	Task involves using both 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 when given support and appropriate formats
	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: The amount of evidence provided can make the question easier or harder. If irrelevant information is provided, students will have to sort to find the appropriate evidence to use. If relevant information is provided, finding evidence to support a claim will be easier.

As illustrated in Table 1, design patterns provide 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, other KSAs involved include the ability to analyze and interpret data to back up an explanation or the ability to view a given situation from a scientific perspective. Next, the design pattern lays out potential observations that would provide evidence that a student possessed the KSAs listed above. In addition, the design pattern lists characteristic features of a task that would elicit the observations needed as proof 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 (Lee and Songer, 2003), observations we may look for would be that the claim represents an understanding of the evidence or data given and that students use appropriate and sufficient data to back up their claim. Tasks that would allow for these observations, such as an open-ended written response item or spoken explanation of a situation, are the kinds of tasks that our assessments would need to employ to gather information about students' abilities to formulate explanations using evidence.

Content-Inquiry Matrix Matched to Each Design Pattern

Although the tasks formulated around a single design pattern will have certain features in common, not all tasks will look 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, 2002). The tasks focused around each design pattern can vary in terms of format, content focus, and complexity. The variable features section of the design pattern lays out some of the ways in which to vary the difficulty of the task. It is important to assess inquiry at various levels so we can evaluate both current levels of reasoning relative to a particular design pattern, and the change or development of complex reasoning over time.

In order to systematically map students' developing knowledge trajectories relative to a particular design pattern (which represents a particular dimension of inquiry reasoning), we created a content-inquiry matrix that classified all assessment tasks with regard to two salient dimensions. First, we examined the amount of content understanding that a student needed to have in order to perform the task. Some questions require very little content knowledge (simple), while others require an in-depth understanding of the content (complex). Second, tasks can vary in the level of guidance provided for inquiry reasoning. In other words, some tasks provide a great deal of support to guide students in the development of a claim and relevant evidence (Step 1). This is often the case for multiple choice items where the student is asked, for example, to match a given claim to one of a set of explanations provided in the multiple choice responses. In contrast, other tasks require students to construct claims and explanations without and guidance or prompting (Step 3). These two dimensions capture the different levels of content and inquiry knowledge needed for a set of tasks that address the same design pattern. Table 2 presents the Content-Inquiry Matrix for the design pattern, "formulating scientific explanations from evidence." Shaded cells indicate levels corresponding to the highest number of corresponding assessment tasks (e.g. step 1 simple, step 2 moderate, step 3 complex tasks are most common).

Our content-inquiry matrix builds from Baxter and Glaser's 'Content-Process Space of Assessment Tasks'(Baxter and Glaser, 1998), which is used to describe characteristics of performance assessment tasks. Baxter and Glaser identify four quadrants that performance assessment tasks can fall into based on the amount of content (content-rich or content-lean) and the amount of freedom students are given with regards to process skills (constrained or open). Our matrix also looks at the amount and type of content required. In addition to the amount of content, we found it important to also 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, whereas other forms of content knowledge require that students understand scientific processes 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 (design pattern) and, in turn, outlines in explicit detail the potential observations and characteristic features associated with each task in a given cell of the table, and give values to the variable features as listed in the design pattern. Baxter and Glaser's quadrants are more generalized for all types of inquiry skills and can be used to classify inquiry assessment questions, but their resources are not meant to give specific guidance in the creation of particular inquiry tasks.

Assessment Design

Using one design pattern and the corresponding inquiry-content matrix, coordinated sets of assessment tasks were developed using both a reverse and forward design process. The reverse design process consisting of mapping existing assessment items to existing design patterns and matrix cells, including the design pattern for "formulating scientific explanations using evidence". Although some items map to particular matrix cells, we did not have a complete set of assessment tasks at the end of the reverse design process. Therefore, the design pattern and matrix was also used to forward design new tasks associated with aspects of inquiry that aligned with our particular curricular learning goals.

Table 2: Levels of content and inquiry for tasks focused on "formulating scientific explanations from evidence"

Inquiry Level	Content Complexity		
	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.

In mapping old tasks and creating new tasks, we used the content-inquiry matrix to make sure that we were examining a range of levels of content and inquiry knowledge. Most of the tasks fell into one of three categories, either a Step 1 simple; Step 2 moderate; or Step 3, complex. In examining the other boxes, 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, the development of tasks to evaluate more complex inquiry, e.g. Step 2 or Step 3 tasks, was much more challenging. This was especially true when we intended to keep the level of content knowledge required relatively low (e.g. simple). The realization of the complexity of developing tasks that were low in content knowledge but high in inquiry reasoning is congruent with our belief that inquiry skills are directly linked to content understandings, and that, particularly at higher inquiry levels, it is difficult to tease apart content development from inquiry skills development. It may be possible to have some level of basic inquiry reasoning skills without fully grasping the content knowledge; however, when practicing inquiry at higher levels, the content is so infused with the inquiry practices that it is difficult to separate the two. Thus, we did not attempt to create any tasks that were high in content knowledge but low in inquiry reasoning (e.g. step 3 simple tasks). For more information on the design and redesign of assessment tasks to match PADI assessment structures see Songer and Wenk, 2003.

Longitudinal Research Associated with Cohorts of Students

As a part of the LeTUS (the Center for Learning Technologies in Urban Schools) partnership in Detroit Public Schools, we have begun to follow cohorts of students as they participate in multiple inquiry-based science curricula coordinated to foster inquiry reasoning skills focused around particular design patterns, e.g. "formulating scientific explanations from evidence". Our work to asses students' development of particular complex reasoning skills over multiple units involves a transition from the design of assessments to measure content and inquiry before and after a single science unit to the design of assessment systems to evaluate complex reasoning in one, two, or up to six coordinated inquiry units. As stated earlier, we believe that the development of inquiry reasoning takes time, and such development is mostly likely an incremental process that may be poorly measured by current assessment

approaches. However, participation and evaluation throughout several carefully sequenced science units may realize a learning trajectory that provides more sensitive information about the development of complex reasoning skills over time and topic. In order to have the ability to measure student knowledge development within and across programs, we are currently forward and reverse engineering assessment tasks to evaluate cohorts' performance on the same design patterns expressed in sequential curricular units.

Our work recognizes that different science disciplines have very different content and different representations of scientific information and data. Design patterns are neutral to content so they can be used with a set of curricula units (Mislevy 2002). However, we are not stating that inquiry skills are content neutral.

> "It is important to emphasize that by having design patterns that are applicable across different content areas, we are not implying that inquiry should be considered a set of generalized skills that can be assessed in the absence of science content. Instead, the goal is to create design patterns that can be instantiated in a wide variety of science disciplines…" (Mislevy 2002).

Thus, while our assessment tasks in each of the different content areas use different representations of evidence and different contexts for reasoning, several tasks have been created using the same design pattern at each of the equivalent content-inquiry matrix levels. Results from longitudinal studies to evaluate students' development of complex reasoning associated with particular inquiry skills and design patterns are forthcoming.

Student Learning Outcomes Associated With One Design Pattern

During the first year of research discussed in this paper, we investigated the kinds of evidence possible with three different categories of assessment tasks: 1) released standardized test items designed to evaluate students' content understandings relative to this unit, 2) released standardized test items and forward-designed tasks designed to evaluate students' inquiry reasoning within the content area of this unit, and 3) reverse and forward-engineered tasks designed to specifically evaluate students content-inquiry reasoning associated with the particular inquiry reasoning skill, "formulating scientific explanations from evidence". This paper will present the results obtained from analysis of performance on these three categories of assessment tasks.

A cohort of 163 primarily 5th graders in five inner city schools representing 94% underrepresented minorities served as the sample for this study. Students' learning outcomes associated with an eight week, inquiry-fostering biodiversity unit were determined before and after the curricular intervention. Assessment consisted of a total of 51 points. The multiple choice instrument consisted of a total of 20 released standardized test and forward engineered task items designed to evaluate students' understanding of science content. Tasks corresponded to the major content of the unit, including the science concepts of biodiversity, food webs, animal classification, and species interactions, as well as inquiry reasoning associated with low complexity levels (e.g. step 1 simple) of the content-inquiry matrix corresponding to the design pattern, "formulating scientific explanations from evidence".

The open-ended instrument consisted of tasks representing 31 points, and largely evaluated students' reasoning with less scaffolded and more complex levels of scientific content and inquiry reasoning (e.g. step 2 moderate and step 3 complex problems). Figure 1 presents a sample step 3, complex task from the open-ended test where students needed to develop a claim and provide justifying evidence to support the claim to explain what would happen to the algae in the pond when the small fish died. A total of 15 assessment points mapped to this design pattern, including 9 points at the step 3 complex level, 5 points at the step 2 moderate level, and one item at the step 1 simple level. For more information on assessment tasks mapping see (Songer and Wenk 2003).

Student learning results were determined both before and after the eight-week biodiversity curricular intervention. Figure 2 left side presents students' pretest and posttest performance on the multiple-choice test (content and low complexity inquiry levels) and the open-ended test (complex reasoning and high inquiry complexity levels). Figure 2 right side presents pretest and posttest performance on tasks corresponding to the design pattern, "formulating scientific explanation from evidence" at each of the three major complexity levels (step 1 simple, step 2 moderate, and step 3 complex). Results from Figure 2 left side demonstrate students' significant pre/post gains on the multiple choice and open-ended instruments. Results from Figure 2 right side demonstrate significant gains on and all three complexity levels associated with "formulating scientific explanations from evidence". Overall scores are higher on simple and more scaffolded tasks (e.g. Step 1 simple), and scores are lower on complex, unscaffolded tasks (e.g. Step 3 complex). Note that despite the complex, unscaffolded nature of the step 3 complex problems, 41% of 5th/6th graders in our inner city schools successfully created unguided claims and evidence relative to the highest science and inquiry levels associated with biodiversity science content.
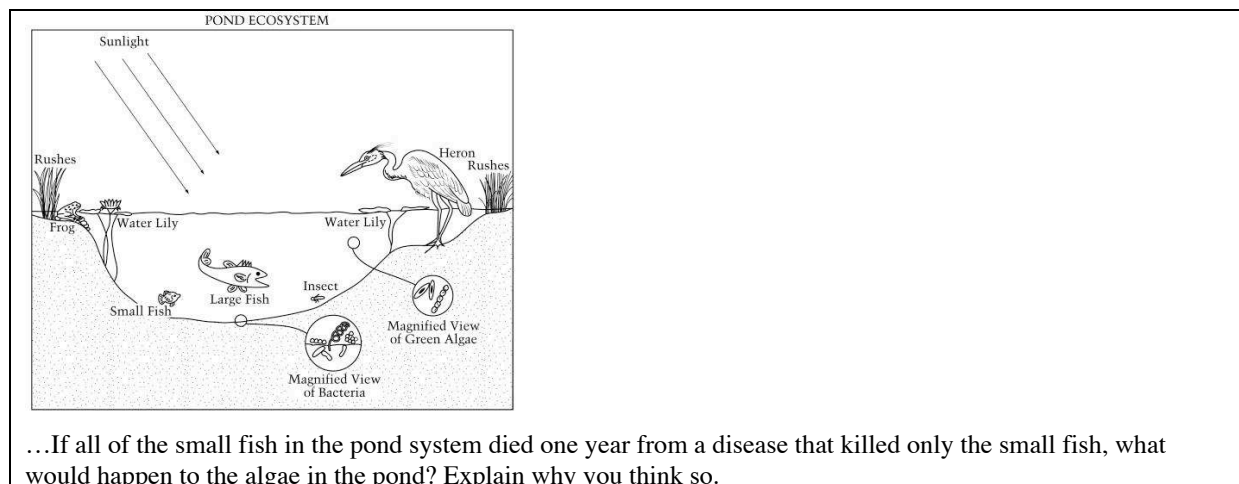
Figure 1: Open-ended assessment task evaluated at step3, complex level relative to the design pattern, "formulating scientific explanations using evidence"
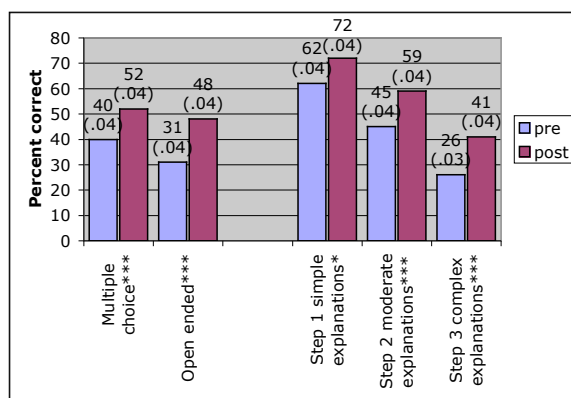

Figure 2: Student learning outcomes on assessment instruments measuring multiple choice, open-ended, and complex reasoning associated with the design pattern "formulating explanations from evidence". (Numbers in parentheses indicate standard error).

Discussion

While educational research in the learning sciences has often utilized pre/post gains on multiple choice, open-ended, and off-the-shelf instruments as convincing evidence of knowledge development, recent discussions about the need for evidence-based research in education (e.g. Shavelson and Towne, 2002) as well as acknowledgements of the mismatch between standards-based reforms and high-stakes tests have led to a need for learning sciences research that expands our thinking about appropriate forms of evidence to evaluate complex reasoning in science. In our own research over the past several years, we have utilized pre/post gains on a range of assessment instruments to demonstrate students' learning associated with our own curricular interventions (e.g. Songer, Lee and McDonald, 2003). Recent examinations of our own and others' assessment approaches indicate that this type of information is not robust enough to provide concrete evidence of complex reasoning in science for several reasons. First, while evidence such as ours in Figure 2 demonstrates significant pre/post gains on multiple choice and open-ended tests, these instruments provide only rough estimations about the kinds of difficulties students are encountering in developing an understanding of what it means to do inquiry reasoning. Without tasks specifically mapped to design patterns such as "formulating scientific explanations from evidence", we were unable to pinpoint the specific kind of inquiry reasoning students were not demonstrating. Second, our previous pre/post results indicated what particular questions students had difficulty with, but without mapping of tasks to complexity levels such as available through our content-inquiry matrix, we were not able to characterize intermediate levels of

knowledge development around complex reasoning such as formulating explanations. Content-inquiry matrices coupled with learning outcomes at each level demonstrate performances at simple, moderate, and complex levels of reasoning associated with the same inquiry skill. Such coordination between tasks and underlying assessment systems allow the possibility of providing more detailed analysis and subsequent intervention in both science content and the level of inquiry knowledge demonstrated. Third, our own research and that of others has not attempted to develop an underlying conceptual framework for systematically assessing complex reasoning matched specifically to learning goals and assessment tasks. While still in the early stages of realizing research outcomes, such an approach has the potential to provide comprehensive clusters of matched tasks, learning goals, design patterns and other dimensions of the design framework, and systematic outcomes from thousands of students. These data can then be used to systematically provide researchers and key stakeholders with concrete evidence to complement high-stakes tests. Research data collected in 2003-4 will result in quasi-experimental comparisons between control and experimental classes relative to students' complex reasoning with several aspects of complex thinking in science.

In summary, research to expand and evaluate new systems for measuring students' complex reasoning in science has been developed, and results are emerging. Assessment systems that coordinate standards-based educational goals, assessment tasks, and the technical specifications of a comprehensive assessment system provide one research-based alternative to the evaluation of curricular reforms leading to evidence of student learning associated with particular dimensions of inquiry reasoning.

References

Baxter, G. P., and Glaser, R. (1998). Investigating the cognitive complexity of science assessments. Educational Measurement: Research and Practice 17(3), 37-45.

Bransford, J. D., A. L. Brown, et al., Eds. (2000). How People Learn: Brain, Mind, Experience and School. Washington D.C., National Academy Press.

Huber, A., Songer, N.B. and Lee, S.Y. (2003). BioKIDS: A Curricular Approach to Teaching Biodiversity through Inquiry in Technology-Rich Environments. Paper presented at the meeting of the National Association of Research in Science Teaching, Philadelphia, PA.

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

Linn, M. C., Lewis, C., Tshuida, I., and Songer, N.B. (2000). Beyond fourth-grade science: Why do U.S. and Japanese students diverge? Educational Researcher 29(3), 4-14.

Mislevy, B., and the Principled Assessment Design for Inquiry Project (in alphabetical order: Chudowsky, N., Draney, K., Fried, R., Haertel, G., Hamel, L., Kennedy, C., Long, K., Morrison, M., Pena, P., Rosenquist, A., Songer, N., Wenk, A.) (2002). Design Patterns for Assessing Science Inquiry. Menlo Park, CA: SRI International.

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

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

National Research Council (2000). Inquiry and the National Science Education Standards. Washington, D.C., National Research Council.

National Research Council (2001). Classroom assessment and the National Science Education Standards. Committee on Classroom Assessment and the National Science Education Standards. Center for Education, Division of Behavioral and Social Sciences and Education. J. M. Atkin, Black, P.l., and Coffey, J. (Eds.). Washington, D.C., National Academy Press.

Parr, C., Jones, T., and Songer, N.B. (in press). Evolution of a handheld data collection interface for science learning. Journal of Science Education and Technology.

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

Shavelson, R., and Towne, L. (2002). Scientific Research in Education. Washington, D.C., National Academy Press.

Songer, N. B. (2003). Urban Students' Complex Reasoning in Science. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Songer, N. B., H.S. Lee, and McDonald, S. (2002). Technology-rich inquiry science in urban classrooms: What are the barriers to inquiry pedagogy? Journal of Research in Science Teaching 39 (2), 128-150.

Songer, N. B. and Wenk. A. (2003). Measuring the Development of Complex Reasoning in Science. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Vygotsky, L. (1978). Mind in Society: The Development of Higher Psychological Processes. Cambridge, MA, Harvard University Press.