Measuring the Development of Complex Reasoning in Science

Nancy B. Songer and Amelia Wenk
The University of Michigan
songer@Umich.edu

Paper presented at the American Educational Research Association (AERA)
Annual Meeting, Chicago, IL. Session 69.027
April 25, 2003

Measuring the Development of Complex Reasoning in Science

Abstract

While many science curricular interventions focus on fostering inquiry in four-six week units, much research in science education and the learning sciences suggests that the development of complex reasoning in science (e.g. inquiry thinking) requires much longer amounts of time. Inquiry knowledge development involves both the development of underlying science concepts and the coordinated development of reasoning skills in that context, such as building explanations from data or evidence (NRC, 2000). Interestingly, few research programs coordinate the design of inquiry curricula with inquiry assessment systems that evaluate students' inquiry knowledge development over multiple topics. This paper outlines the description and results from a coordinated assessment system and set of curricular units promoting and evaluating scientific inquiry development from 5-8th grades. The assessment system is build on a template of design patterns that characterize assessment tasks associated with the same inquiry understanding along dimensions of inquiry thinking and content knowledge complexity. Results demonstrate significant student gains in both content and inquiry knowledge development associated with concepts in biodiversity and the design pattern, "formulating scientific explanations from evidence".

Introduction

Perhaps never before has the issue of measurement of student learning in science been so complex and important. International tests demonstrate that American middle school students' achievement in science declines relative to their peers internationally (e.g. Linn, Tsuchida, Lewis, and Songer, 2000), while education reform laws such as the No Child Left Behind Act support higher levels of accountability and larger consequences for poor performance in science by the 2007-08 academic year. 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. The National Research Council (2001) recommends robust assessment instruments that compliment these standards-based curricular programs with parallel goals focusing on measurement of complex reasoning in science.

> "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 in science, 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 students' inquiry thinking such as the ability to develop explanations from scientific evidence. As high-stakes tests often attempt to match the learning goals of the standards-based reform programs but often fall short, schools are placed in an incredibly difficult mismatch between the emphasis of the high-stakes tests and the emphasis of the reform-based programs. What is needed is a systematic assessment program that reflects the thinking and learning outcomes of standards-based reform programs more directly, and that, if possible, is built on design principles intimately aligned with the reform programs themselves.

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, both to provide tangible evidence of student learning trajectories, as well as evidence to evaluate effective reform programs from those that are less effective. The focus of this paper is the description and preliminary results of one systematic assessment program applied to the assessment of student learning in 4-8th grades affiliated with a reform-based curricular program implemented in a high-poverty urban district.

Longitudinal Evaluation of Complex 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 and persistence, and mediation often in the form of scaffolds to guide children's attention to salient features amidst many complexities within natural world reasoning situations (Lee and Songer, in press; Bransford et al, 1999; Vygotsky, 1978). The development of complex inquiry thinking 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 (NRC, 2000). Such development of complex thinking takes time, and is not well suited to short-term curricular interventions. Ideally, children's inquiry knowledge development occurs systematically over multiple coordinated units, programs and years.

Interestingly, few research programs are designed to evaluate students' inquiry knowledge development over multiple programs, units or years. An idealized longitudinal inquiry assessment program would be matched to a coordinated set of inquiry-focused curricula in terms of learning goals, both science content and inquiry thinking goals. In this assessment program, systematicity is necessary both in the development of the coordinated items measuring complex thinking across units, and in the underlying conceptual framework. This idealized assessment program would take into account both the complexities involved in measuring the development of students' knowledge within a scientific discipline (e.g. ecology, weather) as well as the complexities involved in measuring the development of students' inquiry thinking within that discipline, e.g. the reasoning skills associated with interpreting species data or building explanations from atmospheric science evidence. The larger set of end products would include a coordinated set of science activities that foster deep conceptual understandings in a range of science topics, as well as coordinated assessment instruments that measure the development of content and reasoning skills in science. With this systematic approach to curricular and assessment design, it becomes possible for researchers to provide longitudinal trajectories of students' developing understandings of science leading to a much clear view of both what students learn in standards-based reforms, and where the development of complex reasoning falls short of the ideals.

Design Patterns for Measuring Scientific Inquiry

The focus of the Principled Assessment Design for Inquiry (PADI) project is the development of a conceptual framework and a delivery system support structure for the systematic development and implementation of assessment tasks associated with measuring scientific inquiry. Design patterns represent the foundational knowledge unit that comprises the assessment conceptual framework (Mislevy, et al, 2002), the scientific inquiry knowledge represented in assessment tasks, as well as the structure of a coherent assessment argument comprising the conceptual framework. Design patterns fuse the science education/learning sciences work in the design of inquiry assessments (e.g. Jeong, Songer and Lee, in press) with the evidence-centered assessment design framework of evaluation experts (Mislevy, Steinberg and Almond, in press). The design pattern discussed in this paper is "formulating scientific explanation from evidence", and is represented in Figure 1.

The BioKIDS Project and Inquiry Assessment

In BioKIDS: Kids' Inquiry of Diverse Species, curricular units are developed to specifically foster inquiry thinking among 5-8th graders in topics such as biodiversity, weather, and motion. A particular focus of BioKIDS is the development of a 5th grade unit focusing on biodiversity concepts that will serve as the first of several coordinated, inquiry-fostering curricular units. In this eight-week unit, particular inquiry thinking skills such as the development of explanations from evidence are fostered through a carefully scaffolded activity sequence (Songer, 2000; Huber, Songer and Lee, 2003). The development of sensitive inquiry assessment instruments are a central focus of the project, including assessment tasks to evaluate both students' early and developing understandings of biodiversity concepts, and students' early and developing understandings of inquiry thinking.

Figure 1: Design Matrix 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.	
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	
	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.

One characteristic of the curricular sequence leading to the scaffolding of inquiry and content development is the repeated presence of guided-learning approaches. For example, a central science concept fostered in BioKIDS is an understanding of the concept of biodiversity, a definition of which involves several factors on which scientists often disagree. In the BioKIDS program, fifth grade students are asked to collect animal species data on a particular area or the schoolyard in preparation for the development of a claim and evidence addressing the question, "Which zone in the schoolyard has the greatest biodiversity?" Scientists might evaluate which zone is most diverse using Simpson's index, $D = 1\text{-}E\ (n/N)^2$, a formula that represents species evenness taking into account both the total number of animals (abundance) and the number of different species (richness). While our fifth graders are not taught to use Simpson's index, our program does encourage students to develop a qualitative understanding of biodiversity that takes into account species abundance and richness. In order to gain this understanding, students work with the concepts of abundance and richness in complimentary ways throughout several activities, and the repeated presence of approaches makes this challenging concept understandable to students. Similarly, a central inquiry concept emphasized is "building explanations from evidence". As with the biodiversity concept, fifth graders are provided with repeated opportunities to make claims, determine what evidence is salient, and build explanations from data towards a deep understanding of inquiry thinking with biodiversity concepts.

Assessment Design in Biodiversity

Both a reverse design process and forward design process were utilized for the development of inquiry tasks associated with the fifth grade BioKIDS curricular program. In some cases, existing BioKIDS assessment items were mapped to existing design patterns, resulting in reverse design. Forward design consisted of the development of new tasks and design patterns coordinated with each other. One central inquiry understanding not previously represented in the existing design patterns was "formulating explanations from evidence". In the curricula, this inquiry skill is scaffolded several times throughout the curriculum. This design pattern and the coordinated curricular activities were developed in concert to emphasize the connection between scaffolding and evaluation of "formulating scientific explanations using evidence." Other design patterns that were especially salient to our curriculum were "interpreting data," "re-expressing data," and "using the tools of science."

When mapping our existing items to particular design patterns, there was a wide range of inquiry tasks falling within a single design pattern. These tasks varied in two different dimensions. First, tasks varied in the amount of content understanding that a student needed to have in order to perform the task. Some questions required very little content knowledge (simple), while others required an in depth understanding of the content (complex). Second, tasks varied in the level of inquiry skill that the student needed in order to perform the task. Some of the tasks required a basic level of inquiry ability (Step 1) especially the multiple choice items, while others required students to construct complex answers (Step 3). We developed a matrix that mapped each design pattern along two dimensions capturing the different levels of content and inquiry knowledge needed to answer the questions.

Distinguishing between the different levels of content required that we look at not just the amount of content required to answer a questions, but also look at the type of content knowledge required to answer the question. For example, some content knowledge involved 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.

Each level up (simple to moderate to complex) requires more (quantity) and more difficult content knowledge. A simple question provides all of the content information that a student requires in order to complete the task successfully. A moderate question still gives students information, however, students must either interpret this information to make sense of it in relationship to the question, or the student must bring in other content knowledge that is not supplied in the question in order to successfully complete the moderate task. A complex task requires even more content knowledge and/or interpretation. For a complex question, students are often provided with evidence that is irrelevant to the question, and they must be able to distinguish between relevant and irrelevant information.

An illustration of the design pattern "formulating scientific explanations from evidence" provides examples of the different inquiry abilities required in different tasks. In this design pattern, a Step 1 classification refers to a task where the student is given both the claim and the evidence and they must match the relevant evidence to the given claim. A Step 2 classification refers to a task that requires the student to choose from a list of claims and then construct an explanation using the given evidence. This process is scaffolded through sentence starter prompts and by giving students a choice of possible claims. A step 3 classification of a task requires students to construct their own claim and explanation using the given evidence without any scaffolds or prompts. Table 1 illustrates the different levels of knowledge needed to perform each type of task.

Table 1: 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.

When we had finished creating this table, we re-classified each of our previous assessment items according to the matrix. Each task was classified along both content knowledge and level of inquiry ability needed to perform each task.

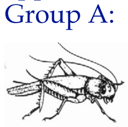
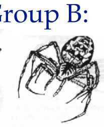
When looking at the placement of tasks along this matrix, patterns emerged. Most of the tasks fell into one of three categories, either a Step 1 simple; Step 2 moderate; or Step 3, complex. While some tasks fell into other boxes on the matrix, the shaded boxes on the diagonal were the most heavily populated. Table 2 below gives examples of tasks that fell into each of the shaded areas.

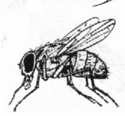
Continuing with forward design of new tasks, interesting patterns emerged. First, the development of Step 1 simple, moderate, or complex tasks were relatively easy to generate. In contrast, the development of tasks to evaluate more complex inquiry, e.g. Step 2 or Step 3 tasks, were much more challenging, particularly if we were intending to keep the level of content knowledge relatively low (e.g. simple). This realization 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 basic inquiry skills without fully grasping the content knowledge; however, when practicing inquiry at higher levels, the content is so infused with the inquiry practices, it is difficult to separate the two.

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

<table>
<tr><th>Question</th><th>Step and Complexity Level</th></tr>
<tr><td>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</td><td>Step 1, Simple</td></tr>
<tr><td>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)</td><td>Step 2, Moderate</td></tr>
<tr><td>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.

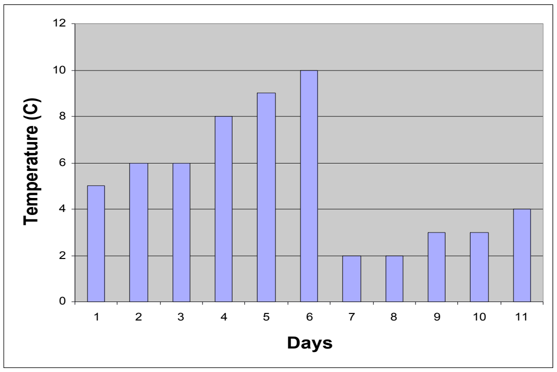
What would happen to the large fish? Explain why you think so.</td><td>Step 3, Complex</td></tr>
</table>

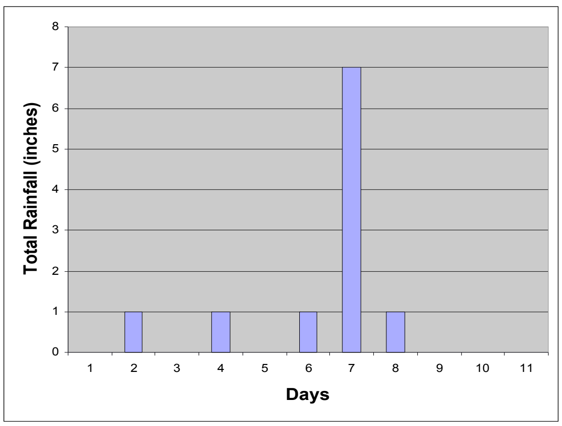
Assessment Design for Additional Inquiry Units: Weather and Motion

Our project is also interested in assessing inquiry ability over time. As stated above, participation in a single inquiry-based science curriculum may only slightly improve a student's inquiry skills. However, participation in several, carefully sequenced science inquiry curricula may have the desired effect of increasing both students' content and inquiry knowledge. We are beginning to expand our assessment system to determine the kind and nature of students' inquiry development over multiple inquiry-based science curricula, including a sixth grade weather program and a seventh grade motion program. As a part of LeTUS (the center for Learning Technologies in Urban Schools), we will be following cohorts of students as they progress from upper elementary school through middle school as they participate in multiple inquiry-based science curricula. In order to accomplish this task, we have both reverse designed and forward-designed assessment tasks for both the weather and motion programs. To maintain an ability to measure student knowledge development within and across programs, we will evaluate student performance on some of the same design patterns in weather and motion as in the biodiversity curriculum. Table 3 presents two tasks that we created for weather based on the design pattern "formulating scientific explanations from evidence."

Table 3: KGS (Weather) Questions Mapped to the Level of the "Formulating Scientific Explanations Using Evidence" Design Pattern

Question	Step and Complexity Level
1. A meteorologist predicted that the temperature was going to drop in the next few days, and that it would be bright and sunny. Choose the evidence below that would support her claim. a. A cold front is moving ahead of a low pressure system b. A cold front is moving ahead of a high pressure system c. A warm front is moving ahead of a low pressure system d. A warm front is moving ahead of a high pressure system	Step 1, moderate
For the past twelve weeks, Mr. Lee's Science class has been recording the weather conditions. They have recorded their findings in the following two charts:	Step 2, moderate





During which weeks did a cold front **MOST LIKELY** collide with an existing warm air mass?
I think that a cold front collided with an existing warm air mass on day _____ because…
(give two reasons to support your answer)
1.
2.

Student Learning Outcomes on One Design Pattern

Student learning associated with all three content and inquiry complexity levels associated with one design pattern was determined both before and after an eight-week biodiversity curricular intervention. The sample consisted of 163 primarily 5th grade students in five urban schools containing 94% underrepresented minorities. Table 4 illustrates students' learning across three levels of the inquiry complexity and content knowledge for the same design pattern, "formulating scientific explanation from evidence". Data evaluating these students' weather and motion learning across three levels of inquiry complexity and content knowledge are being collected beginning in Spring '03, and will be reported in future papers.

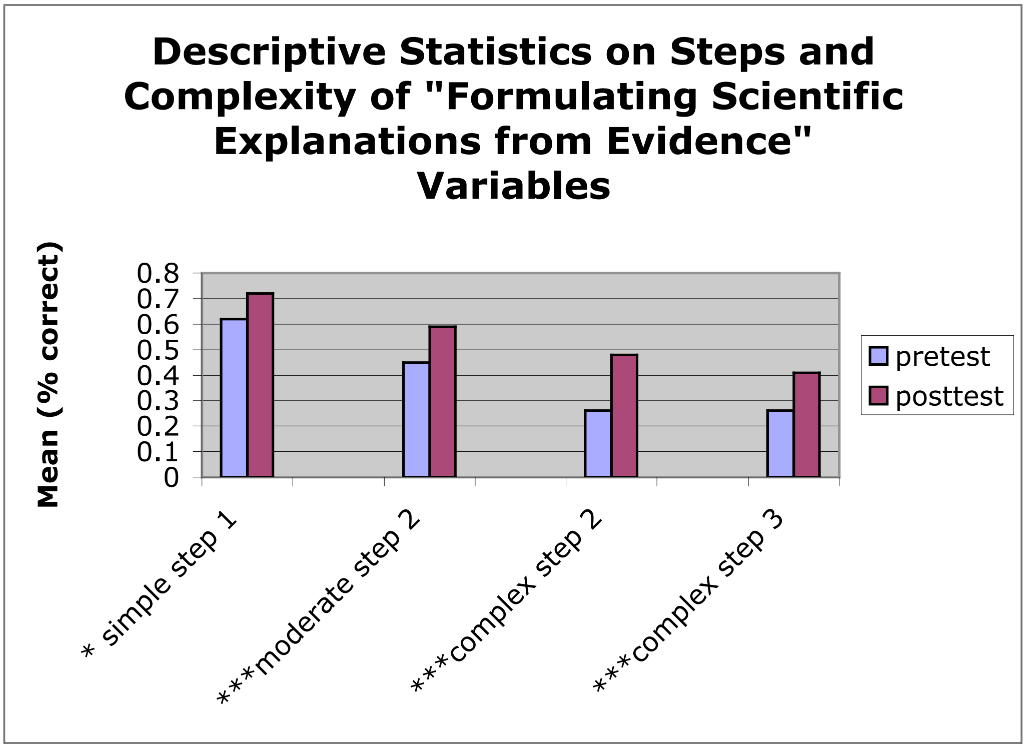


Table 4: Descriptive Statistics on Steps and Complexity of "Formulating Scientific Explanations from Evidence"

As Table 4 illustrates, students exhibited the lowest pre and posttest scores on the tasks characterized as complex. Similarly, students demonstrated lower scores on tasks requiring students to provide unscaffolded inquiry thinking (e.g. Step 3) as opposed to guided inquiry thinking (Step 1). In every case, students made significant gains from pretest to posttest on all of the tasks measuring all levels of steps/complexity variables, showing that students improved both their understanding of science content associated with biodiversity after the intervention, and their understanding of inquiry reasoning after the BioKIDS intervention.

Conclusions

Previous research diagnosing urban sixth graders' scientific inquiry skills revealed that few students recognized what kinds of evidence are relevant to support claims or could differentiate explanations from evidence (Jeong, Songer and Lee, submitted). An implication of this earlier work is that systematic curricular programs are needed that foster and scaffold the development of inquiry thinking such as building explanations from evidence, and that such curricular programs need to be coupled with systematic assessment systems to provide sensitive evaluation of early and intermediate levels of inquiry knowledge development. This paper explains one systematic approach to inquiry assessment developed in conjunction with a multi-year, coordinated inquiry curricular program. Research results demonstrate an assessment system that is sensitive to the development of students' complex understandings of science concepts with biodiversity concepts. Research results also demonstrate an assessment system that is sensitive to the development of students' complex reasoning with complex reasoning. Early results suggest that at the lower levels of conceptual and inquiry development, assessment systems might be able to tease out the development of either content or inquiry thinking. However, at the higher conceptual and reasoning levels, the development of understandings of science content and inquiry thinking are intertwined. This research has implications for the design of curricular scaffolds for fostering conceptual development and inquiry thinking and the design of systematic assessment systems to foster life-long learning in science.

References

Bransford, J. D., Brown, A. L., Cocking, R. R. (1999). 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 annual meeting of the National Association of Research in Science Teaching (NARST), Philadelphia, April.

Lee, H-S. and Songer, N.B. (in press) Making Authentic Science Accessible to Students. *The International Journal of Science Education.*

Jeong, H., Songer, N.B., Lee, S.Y. (submitted) Diagnosing Urban Sixth Graders' Inquiry Skills and Understanding. *Science Education.*

Linn, M.C., Tsuchida, I., Lewis, C. and Songer, N.B. (2000) Beyond Fourth Grade Science: Why do US 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. Technical Report, SRI International.

Mislevy, R., Steinberg, and Almond, (in press) On the structure of educational assessments. To appear in *Measurement.*

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

National Research Council (2000) Inquiry and the National Science Education Standards. Committee on Development of an Addendum to the National Science Education Standards on Scientific Inquiry. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.

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

Vygotsky, L. (1978) Mind in Society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.