JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 49, NO. 2, PP. 141–165 (2012)

Guiding Explanation Construction by Children at the Entry Points of Learning Progressions

Nancy Butler Songer[1] and Amelia Wenk Gotwals[2]

[1]*School of Education, University of Michigan, Ann Arbor, Michigan*
[2]*College of Education, Michigan State University, East Lansing, Michigan*

Received 7 June 2011; Accepted 16 November 2011

Abstract: Policy documents in science education suggest that even at the earliest years of formal schooling, students are capable of constructing scientific explanations about focal content. Nonetheless, few research studies provide insights into how to effectively provide scaffolds appropriate for late elementary-age students' fruitful creation of scientific explanations. This article describes two research studies to address the question, what makes explanation construction difficult for elementary students? The studies were conducted in urban fourth, fifth, and sixth grade classrooms where students were learning science through curricular units that contained 8 weeks of scaffold-rich activities focused on explanation construction. The first study focused on the kind and amount of information scaffold-rich assessments provided about young students' abilities to construct explanations under a range of scaffold conditions. Results demonstrated that fifth and sixth grade tests provided strong information about a range of students' abilities to construct explanations under a range of supported conditions. On balance, the fourth grade test did not provide as much information, nor was this test curricular-sensitive. The second study provided information on pre–post test achievement relative to the amount of curricular intervention utilized over the 8-week time period with each cohort. Results demonstrated that when taking the amount of the intervention into account, there were strong learning gains in all three grade-level cohorts. In conjunction with the pre–post study, a type-of-error analysis was conducted to better understand the nature of errors among younger students. This analysis revealed that our youngest students generated the most incomplete responses and struggled in particular ways with generating valid evidence. Conclusions emphasize the synergistic value of research studies on scaffold-rich assessments, curricular scaffolds, and teacher guidance toward a more complete understanding of how to support young students' explanation construction. © 2012 Wiley Periodicals, Inc. J Res Sci Teach 49: 141–165, 2012

Keywords: learning progressions; assessment; elementary

One of the new priorities of American policy documents is the increasing importance of the development of 21st century learning skills (Partnership for 21st Century Skills, 2009). Some of these skills include global awareness, environmental literacy, creativity, and critical thinking and problem solving (Partnership for 21st Century Skills, 2009). Several of these skills, such as creativity and global awareness, are not well represented in discipline-based national standards such as the science standards (e.g., National Research Council, 1996). On balance, recent policy documents (e.g., NRC, 2007) and standards-related documents (The College Board, 2009; NRC, 2011) place strong emphasis on one of these skills—critical

Correspondence to: N.B. Songer; E-mail: songer@umich.edu
DOI 10.1002/tea.20454
Published online 9 January 2012 in Wiley Online Library (wileyonlinelibrary.com).

thinking and problem solving—in the form of explanation building about focal science content.

An emphasis on explanation building about core ideas in science is not new, as this idea was, for example, promoted in the 1996 National Science Standards (NRC, 1996). This forward thinking document and its companions (e.g., NRC, 2000) emphasized inquiry skills such as explanation building and provided strong narrative examples of elementary students' successes in building and critiquing explanations, such as a narrative illustrating fifth graders' quest to build explanations for why three trees in their schoolyard were growing differently in different locations (NRC, 2000; p. 6). Table 1 highlights the presentation of sample standards from the content (e.g., organisms) and inquiry sections (e.g., explanation) of the 1996 national science standards document. Table 1 also illustrates similar content and science practices

Table 1
Contrasting presentation of explanation standards, content standard, and their fusion from 1996 and 2009 standards documents

	Explanation Standard	Content Standard	Fusion Explanation With Content
National Science Education Standards (1996)	"Develop descriptions, explanations, predictions and models using evidence" "Students should base their explanation on what they observed, and as they develop cognitive skills, they should be able to differentiate explanation from description—providing causes for effects and establishing relationships based on evidence and logical argument" (Content Standards 5-8, p. 145).	"Populations and Ecosystems" "The number of organisms an ecosystem can support depends on the resources available and abiotic factors, such as quantity of light and water, range of temperatures, and soil composition . . . Lake of resources and other factors, such as predation and climate, limit the growth of populations in specific niches in the ecosystem" (Content Standards 5-8, p. 158).	Separate lists of explanation and content standards. No fusion explanation with content. Rich narratives with no information about how to get there.
Science: College Board Standards for College Success (2009)	"SP 4.1 Constructing Explanations" "Students construct explanations that are based on observations and measurements of the world, on empirical evidence and on reasoning grounded in the theories, principles, and concepts of the discipline" (Science Practices, p. 14).	"Objective LS 3.2 Interactions of Living Systems" "Students understand that organisms in all ecosystems interact with and depend on each other, and that organisms with similar needs compete for limited resources" (Middle School Life Science, p. 67).	Separate presentation of explanation and content standards, however fusion is explicit through performance expectations. "Performance Expectation LSM-PE3.2.7 Explain, using information about the needs and behaviors of the invasive species, why invasive species are often able to increase rapidly and why the numbers of other organisms either increase or decrease when invasive species enter an ecosystem (p. 67)."

from a more recent standards document (The College Board, 2009). Notice that while both standards list explanation standards as separate from content standards, the 2009 standards document presents a specific means of fusing explanation and content standards: through performance expectations that blend explanation building and core content knowledge into one. The College Board defines a performance expectation as follows: ''The performance expectations illustrate how students engage in science practices in order to develop a better understanding of the objective and the essential knowledge statements'' (2009; p. xix). On balance, the 1996 document provides no means of how to fuse content with explanation building. In addition, the 1996 document provides no guidelines to inform curriculum developers, teachers or researchers in how to lead students, particularly young students, towards the guided construction of explanations. With thought provoking but idealized illustrations of science inquiry in standards documents of the 1990s, perhaps it is not surprising that in the past two decades, American students have consistently underperformed relative to peers on international assessments that evaluate students' abilities to construct explanations about focal content. In one well-documented international comparison designed to provide indicators of data interpretation and the critique of scientific evidence explanations by 15-year-old students worldwide, American students performed poorly overall including a rank of 29th out of 57 countries, an average that was significantly below the OECD average (OECD, 2007).

Building on Existing Research

Several research groups have specifically focused on fostering explanation building or argumentation with middle or high school audiences, with some notable successes. Some groups have drawn from the work of Toulmin (2006) and others to define explanations and to design and study middle and high school students' explanation building in science (e.g., Linn, Shear, Bell, & Slotta, 1999; McNeill & Krajcik, 2007; Sandoval, 2003; Songer, Kelcey, & Gotwals, 2009). While the research studies are quite plentiful and strong at the middle and high school levels, only a handful of researchers (e.g., Lehrer & Schauble, 2010; Metz, 1991) have conducted research on the creation of scientific explanations by younger students, such as in the elementary grades of K-6. This lack of emphasis on elementary-age students seems surprising in light of the strong language in the standards encouraging explanation building in grades K-4.

> Even at the earliest grades, students should learn what constitutes evidence and judge the merits or strength of the data and information that will be used to make explanations. After students propose an explanation, they will appeal to the knowledge and evidence they obtained to support their explanations (NRC, 1996; Content Standards K-4, p. 122).

Recent policy documents such as Taking Science to School (NRC, 2007) suggests that not only are younger students capable of scientific practices such as explanation building, but that the development of scientific practices such as explanation building should begin well before middle or high school, should be fostered in structured and guided ways, and should be evaluated through high stakes and classroom assessments.

> In short, young children have a broad repertoire of cognitive capabilities directly related to many aspects of scientific practice, and it is problematic to view these as simply a product of cognitive development. Current research indicates that students do not go through general stages of cognitive development, and there are no 'critical periods' for learning particular aspects of science. Rather, cognitive capabilities directly related to

> scientific practice usually do not fully develop in and of themselves apart from instruction, even in older children or adults. These capacities need to be nurtured, sustained, and elaborated in supportive learning environments that provide effective scaffolding and targeted as important through assessment practices. (NRC, 2007; pp. 44-45).

This excerpt emphasizes a few important points, but two are particularly important. First, it is crucial to develop, and consequently study, nurturing in the form of cognitive scaffolds that can guide younger students in successful explanation building. Second, it is important to recognize that developing strong assessment instruments to provide us with feedback on students' successes and failures is an essential dimension of our research work, even if this piece of the work may be particularly challenging.

Collectively, the conclusions from these documents coupled with a shortage of research focused on explanation building by younger, elementary school age students suggests a need for research studies that systematically foster and study students' early experiences with explanation building, such as students in the latter parts of primary grades (e.g., grades 4, 5, and 6 in the USA). In particular, we see a need for research studies focused on the following goals:

(a) Characterization and evaluation of assessment items that are sensitive to elementary students' first attempts at explanation building, and
(b) Characterization and evaluation of the amount, kind and nature of cognitive supports that might be particularly valuable for elementary students' first experiences with explanation building around focal science concepts.

This article addresses these goals through two types of research studies. Our Assessment Study was designed to provide information about how well scaffold-rich assessments evaluated our fourth, fifth, and sixth grade students' explanations under a range of scaffold conditions. This study addressed the research question,

- How well do scaffold-rich assessments, designed in association with a learning progression-framework and associated curricular units, provide information on the development of scientific explanations about biodiversity in late elementary (e.g., fourth, fifth, and sixth grade) populations?

An Achievement and Errors Study focused on pre–post achievement relative to the amount of curricular intervention and information on the types of errors fourth, fifth, and sixth graders demonstrated in explanation construction. This study addressed the research question,

- What learning outcomes and types of errors do late elementary students demonstrate in their first systematic experience with explanation building around focal concepts in biodiversity?

Addressing these research questions required a set of coordinated, sequential steps to produce the necessary products involved in the work. These products included:

(a) The development of learning progressions focused on core ideas in biodiversity and the science practice of constructing science explanations over three consecutive years,
(b) The development of three consecutive curricular units that provided scaffolds for guidance, reflection and repeated exposures to explanation construction around focal concepts throughout the 3-year period, and

(c) The development of scaffold-rich assessment instruments which evaluated the development of explanation construction around focal concepts along the learning progressions.

This article continues with a brief discussion of the development of each of these scholarly products. This is followed by findings about (1) the quality of the learning progression-based assessment instruments and (2) the nature of student learning throughout the implementation of eight week curricular units with fourth, fifth, and sixth grade students.

Learning Progressions Focused on Explanations About Core Ideas in Science

Over the past 5 years, our work has focused on guiding students to construct scientific explanations about focal content in the life sciences over multiple consecutive years. In our work and that of several others, learning progressions are one means of prioritizing focal concepts and guiding the systematic development of core ideas in science across years and grade bands (e.g., NRC, 2011; Songer et al., 2009). We define learning progressions as taking ''a stance about both the nature and the sequence of content and inquiry reasoning skills that students should develop over multiple curricular units and years'' (Songer et al., 2009, p. 612).

We use learning progressions as templates for the coordinated development of our consecutive curricular units, our scaffold-rich assessment instruments and our professional development materials. In our work, the knowledge represented in learning progressions is the knowledge that is most valued. Therefore learning progressions contain both science content and science practice dimensions, and similar to the College Board Standards highlighted in Table 1, there needs to be an explicit means of fusing content with science practices (Songer et al., 2009).

Learning progression development begins in the creation of outlines for content progressions and practices progressions that draw from research results and earlier learning progressions. For this study, we developed a content progression outlining the core science content for our fourth, fifth, and sixth grade units, and a practice progression outlining the scaffolded construction of scientific explanations throughout fourth, fifth, and sixth grades. The development of learning progressions involved a series of discussions among all team members to determine both the most essential explanations about biodiversity and ecology that we intended to anchor our activities and the most logical sequence of these fused explanations about focal content that would satisfy state and local standards. Content and practice progressions and more information about their development are available in Songer et al. (2009).

Once developed, content and practice progressions were used as templates for the development of a series of learning goals that specifically fused sections from each of the progressions together. These learning goals served as the concrete manifestation of the fusion of content and practice that became the anchors for each activity of each curricular unit. Figure 1 presents a sample section from the 3-year content progression, a sample section from the practice progression emphasizing scaffolds for explanation building, and a sample learning goal that served as our means of fusing this content with this practice.

Three Sequential Curricular Units With Written Scaffolds

The second set of products developed were the three sequential curricular units for fourth, fifth, and sixth grades. In order to serve as replacement units within our public school classrooms, each of our curricular units had to address the life science and inquiry standards dictated by the school district and state within the predetermined 8-week period. In addition,

Content progression section: *Ecology 13:* Because many animals rely on each other, a change in the number of one species can affect different members of the web.

Practice progression section: *Ep-Explanation partial scaffolds*: Students build complete scientific explanation consisting of a Claim, two pieces of Evidence and Reasoning with partial (content only) scaffolds.

Learning goal: Students construct scientific explanations to address the question, how have recent changes in the Detroit River affected yellow perch populations?

Figure 1. Sample content and practice progression sections and their fusion.

a central goal was to design the units so that they manifested the fused learning goals into a coordinated set of activities that built systematically on previous activities. As if these challenges were not already overwhelming, a new focus on younger students, our fourth and fifth graders, presented an additional challenge: the design of scaffolds to guide students at the entry points of our learning progressions, in fourth and fifth grades, towards age-appropriate explanation building.

Several crucial areas of literature guided our curricular design. First, we drew on the existing research in explanation construction by elementary age students (e.g., Lehrer & Schauble, 2010; Metz, 1991) and elementary science curriculum materials development (e.g., Davis & Krajcik, 2006) for key insights into age appropriate guidance and scaffolds. We reviewed the literature and learning theories focused on how children learn to be able to do more complicated or ill-structured tasks on their own after guidance from strategic supports (e.g., Vygotsky, 1978). Literature on learning theories and their applications reminded us of the importance of careful guidance, reflection and repeated exposures to material (e.g., NRC, 2007). We also drew from important research on the character of cognitive and procedural scaffolds by others (e.g., Linn, Bell, & Davis, 2004; Quintana et al., 2004; Reiser, 2004). In addition, we revisited earlier work in the development and empirical testing of a system of scaffolds for sixth graders' development of scientific explanations over time and topic (Lee and Songer, 2003; Songer, 2006) and the systematic evaluation of students' fused content and practice in scaffold-rich assessments (Gotwals, 2006; Gotwals & Songer, 2010).

Another driving force in curricular development was the integration of scientific understanding held by the research scientist team members (zoologists), with ideas from teachers and curriculum developer team members who held knowledge about working with students in grades 4–6. The integration of ideas towards the development of curricular units that retained characteristics from the full range of our many authors did not happen without fits and starts. What we have learned about successful work among interdisciplinary science-educator research teams is documented in a forthcoming article (Peters & Songer, forthcoming).

A central dimension of these three curricular units was the guided construction of explanations around focal life science topics aligned with the state and district curricular framework. In our earlier work we were guided by the work of Toulmin (2006) in the development of a sixth grade appropriate definition of explanation that consisted of three parts: claim, evidence, and reasoning (Songer et al., 2009). Figure 2 presents our definition of a scientific explanation as well as our definitions of claim, evidence and reasoning as presented to teachers and students.

Each unit contained a set of activities that culminated in guided explanation construction activities—either the guidance of a claim plus evidence or the guidance of a complete explanation (claim, evidence and reasoning) to address a provided scientific question. Figure 3

Scientific Explanation	A scientific explanation includes a claim, evidence, and reasoning. Scientists use explanations to answer scientific questions.
Claim	A claim is a complete sentence that answers the scientific question.
Evidence	Evidence is observations, data, or information that helps you answer the scientific question.
Reasoning	Reasoning tells why your evidence supports your claim. You can use scientific definitions or ideas to explain why you chose the evidence you did.

Figure 2. Scientific explanation guide.

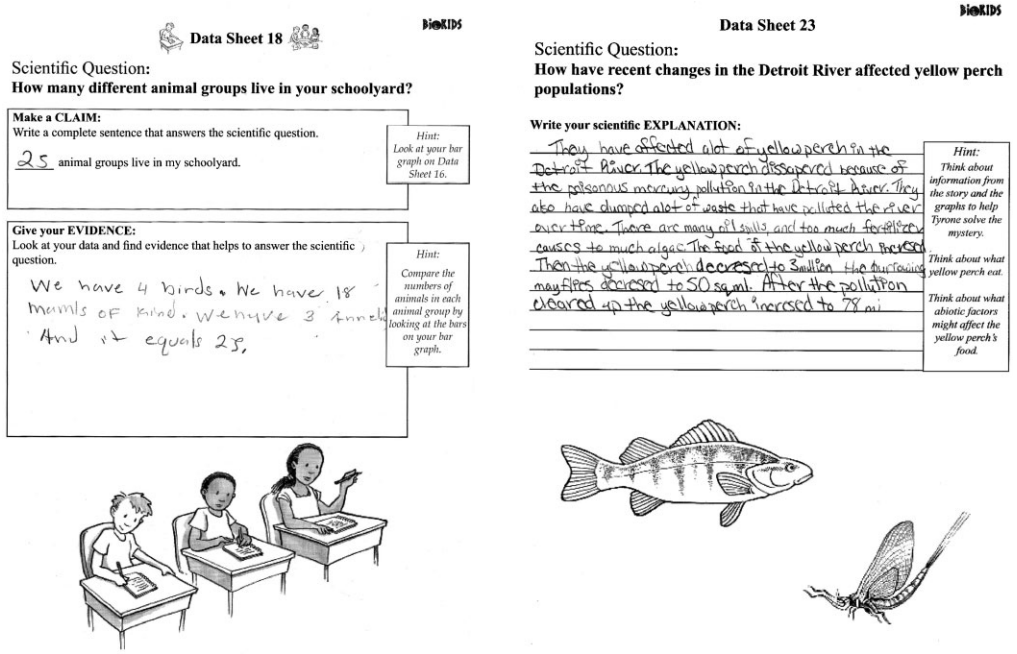

Figure 3. Sample full scaffold activity (e.g., both explanation definitions and content hints) from 4th grade (left) and partial scaffold activity (content hints only, right) from 6th grade. Note: The evidence response in the left example is not correct, as the student is counting the number of animals (abundance), when the question asked for the number of animal groups.

presents two sample pages from student notebooks. The left image is a scaffold-rich claim and evidence activity taken from our fourth grade unit. Note that this activity page contains both science practice scaffolds that define claim and evidence as well as content scaffolds on the right that are specific to this scenario. The image on the right is a sample scaffold-rich explanation page taken from our sixth grade unit that includes only the content hints, as the science practice scaffolds have been faded. This activity is also a manifestation of the learning goal presented in Figure 1. The fourth grade unit had seven guided explanation activities over the 8 weeks, the fifth grade unit had nine and the sixth grade unit had ten.

The process of curriculum development involved several iterative drafts. In each case, activity drafts were developed and revised again through a round of pilot testing with a small set of students and teachers. For the academic year of this study, the curricular units were in their third classroom-based version. In this version, the final units consisted of 18 lessons in fourth grade and 21 lessons in fifth and sixth grade for implementation over an 8-week period (approximately 2.5 lessons per week). Final curriculum units consisted of the following:

(a) Bound student notebooks of approximately 60 pages for each participating student,
(b) Teacher binders of approximately 120 pages of activities (exact copies of student activities plus educative support annotation),
(c) An additional 50 pages of resource material for teachers that included a glossary of key science terms, background information on local ecosystems, information to help in collecting schoolyard data (e.g., a Guide to Invertebrate Identification), information on scientific explanations, and detailed guides for all our technology resources (e.g., how to use the BioKIDS app (Parr et al., 2003) for schoolyard data collection of animals),
(d) Complete class sets of technological tools designed to work with each unit, including BioKIDS app (Parr et al., 2003) on handheld computers for schoolyard data collection, Critter Catalog (Dewey et al., 2011), an online species database of all common animal species, and object-orientated spreadsheets for the organization of student-collected schoolyard data (Figure 3).

Scaffold-Rich Assessments That Fuse Content and Science Practices

Learning progression templates also served as the foundation for assessment development. As advocated in A Framework for K-12 Science Education (NRC, 2011), the evaluation of learning progression products that fuse content and practices requires assessments that similarly fuse content ideas and science practices. Our goal in assessment design was to utilize our learning progressions as templates for the design of three assessment instruments, one matched to the fused content and practices learning goals of the fourth, fifth, and sixth grade curricular units. A driving hypothesis was that each of our assessments needed items focused on the fusion of content with explanation building across a range of scaffolding conditions, some items with more guidance and some with less, to capture the successes and failures students might exhibit in their early attempts at fusing science content with explanation building. In other words, we designed our assessment instruments in line with the nature of learning progressions so that we might gather empirical evidence that was more rich and informative than whether or not students could create a fully coherent scientific explanation.

We developed a total of 27 scaffold-rich assessment items that represented five categories: four categories of explanation-building items with varied levels of difficulty and scaffolds (minimal, intermediate I, intermediate II, and complex), and one category of items focused on the development of a food web, an important area within our content progression.

Minimal items had a high amount of both content and science practice (explanation construction) scaffolds; these items provided students with content-relevant evidence and asked students to match a relevant claim to the provided evidence. Intermediate I items asked students to fully construct an explanation, however the students were provided with hints in both explanation construction and content scaffolds. Intermediate II items also asked students to fully construct an explanation, however the students were provided with only explanation construction hints. In complex items, students constructed an explanation with no scaffolds of any kind. Figure 4 presents a sample minimal item and a sample intermediate I item. For more information on assessment design work, please see Gotwals and Songer (2010) or Gotwals, Songer, and Bullard (in press).

Using this battery of 27 items, we built three baseline tests, Baseline A, B, and C, each of which contained a mixture of items from the content areas of the fourth, fifth, and sixth grade. We used six linking items that were common to all three baseline tests. These items allowed us to calibrate the difficulty of all items relative to each other. We administered these baseline tests to 488 students in the three grades: 188 fourth grade students; 167 fifth grade students, and 133 sixth grade students. Each student took one of the three baseline tests. The students' responses were coded using a four level learning progressions-based coding rubric (for more details see Gotwals et al., in press). Items were calibrated using a partial credit model (Masters, 1982), which extends the Rasch model (Rasch, 1960) by allowing for coding student responses into multiple levels. The data were fit well by the model with all item and student fit statistics falling between 0.75 and 1.25 (which according to Bond & Fox, 2001 indicates good fit to the model).

After calibration of all items, we organized items into separate fourth, fifth, and sixth grade tests that were administered both before and after implementation of each of the curricular units. Table 2 shows the breakdown of items in each baseline and grade-level test. In this round of assessment development, we included three levels of explanation items on the sixth

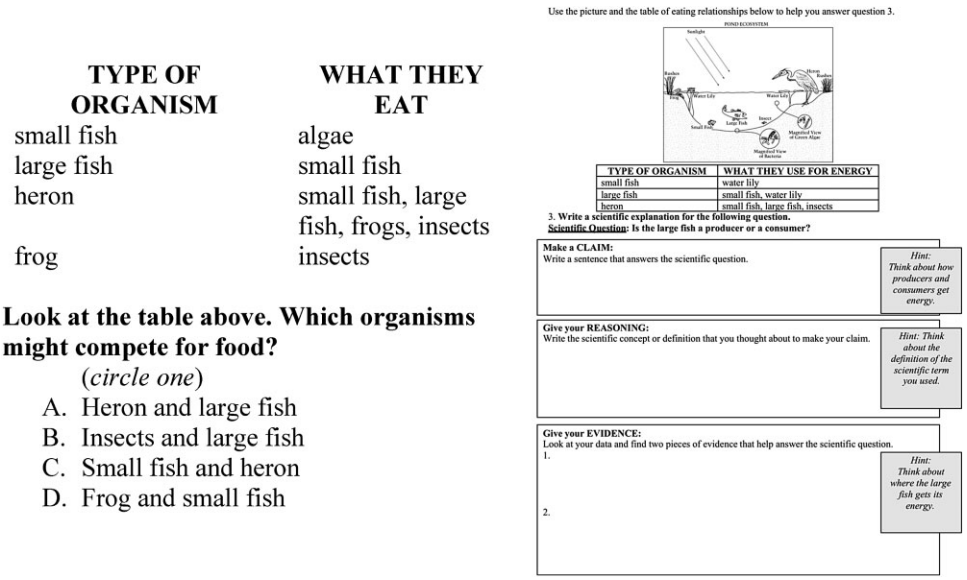

Figure 4. Sample minimal (left) and intermediate I (right) assessment items.

Table 2
Number and type of item on the baseline, fourth, fifth, and sixth grade tests

Item Type	Baseline A	Baseline B	Baseline C	4th Grade	5th Grade	6th Grade
Minimal	3	3	2	2	2	4
Intermediate I	3	3	4	2	2	0
Intermediate II	4	3	3	2	2	2
Complex	2	2	3	2	2	2
Other		1	1	1 (food web)	0	1 (food web)

grade test (we did not include items with both content and structural explanation scaffolds). However, subsequent iterations of development of sixth grade assessments have included a fourth level.

Research Design

Sample

Our sample consisted of 939 students and 23 teachers across 19 research schools within the Detroit Public Schools (DPS). As sixth grade teachers were housed in middle school buildings so that each teacher taught multiple classes of sixth grade science, the number of different sixth grade teachers is smaller than the number of fourth and fifth grade teachers. Our research sample represented approximately 5% of the Detroit Public School fourth, fifth, and sixth graders (19,053 total students in grades 4, 5, and 6). Detroit has a very high poverty rate as compared to state and national averages, with approximately 87% students applying for free or reduced-price meals in 2010. DPS students consistently underperform on state and national standardized tests as compared to state averages (e.g., fifth grade state science passing percentage was 56% in DPS as compared to 82% statewide in 2007). Our sample for the type of errors study consisted of 27 fourth grade and 50 fifth grade notebooks collected from two fourth grade and three fifth grade classes. All students in the type of errors study were subjects from the larger research school samples. As in our previous studies, the students in our research schools were determined to be no more academically advantaged than students in the other schools (e.g., non-research schools) within the Detroit Public School system. For example, our research school students demonstrated similar or lower passing rates on state exams in reading, math and science as students from non-research schools within the Detroit Public School system (Songer et al., 2009). Table 3 presents information about our sample.

Missing Data

Rather than remove students who had missing data, we used a multiple imputation procedure to impute missing values (e.g., Raghunathan, Lepkowski, Van Hoewyk, & Solenberger,

Table 3
Research study population demographics by grade

	Students	Teachers	Ethnicity
4th grade	455	12	84% African American, 10% Hispanic, 6% other/multi-racial
5th grade	294	7	57% African American, 40% Hispanic, 3% other/multi-racial
6th grade	190	4	92% African American, 0.6% as Hispanic and 6.3% as multi-racial
Totals	939	23	86% African American, 12% Hispanic, 2% other/multi-racial

Table 4
Multiple imputation data for fourth, fifth, and sixth grade data

	Pretest Raw Score	Pretest Imputed Score	Posttest Raw Score	Posttest Imputed Score
4th grade ($N = 455$)	6.13 (9.5% missing)	6.05	7.92 (21% missing)	7.95
5th grade ($N = 294$)	6.0 (10% missing)	5.98	9.2 (18% missing)	9.0
6th grade ($N = 190$)	12.34 (11.6%)	12.29	15.14 (15.8% missing)	15.21

2001). At each grade level there were students who were missing either the pretest or the posttest. Because we work in a district with a high level of student mobility and the schools also struggle with absenteeism, this outcome is not surprising. Table 4 presents the averages for pretest and posttest scores before and after multiple imputation for our fourth grade, fifth grade, and sixth grade populations. As the raw and imputed scores are quite similar, the analyses demonstrate that the students who were missing either the pre- or the posttest are likely not very different (in terms of achievement) than students for whom we have all of the data. While we are not able to fully empirically test this (because the data are missing), we utilize these strong results to justify the decision to impute the missing data and use the imputed data for all of our statistical analyses.

Intervention

Our three curricular units were officially adopted by the school district as the life science replacement unit for an 8-week time period in fourth, fifth, and sixth grades. In addition, the district had previously adopted a popular life science textbook series that was approved for this same content material. In this way, teachers could choose to use either our intervention materials or the textbook as a resource for their lessons. As a result, our research studies adopted a dose–response study design (Ruberg, 1989) where the treatment variable was a continuous measure of the percent of the intervention curricular program activities that were completed and recorded in teacher logs. In other words, for every topic of the required life science material, each teacher could choose to use either the district textbook or our intervention curricula (treatment). Due to this study design, it was necessary for us to take into account the amount of intervention used by students as a central dimension of our achievement analyses. In this study, values of the treatment variable ranged from no intervention activities completed (control) to 100% intervention, with averages of 38% (fourth grade, SD = 20%), 47% (fifth grade, SD = 32%), and 61% (sixth grade, SD = 20%).

Regardless of how much of our intervention they used, all teachers in our research schools were provided full support through regular classroom observations and visits by our research staff and monthly professional development workshops focused on the following topics: guiding students in the creation of explanations about focal content, science content in biology, ecology, and biodiversity, discussions of best ways to lead schoolyard data collection, and pedagogical content knowledge associated with our curricular units.

Results: Assessment Study

The first study investigated how well our scaffold-rich assessments, that were designed in association with a learning progression framework and associated curricular units, evaluated the development of explanations about ecology and biodiversity in fourth, fifth, and sixth grade populations. Psychometric analyses were conducted to gather evidence of the amount and character of information our items provided relative to each other, as well as what kinds

of student profiles were well matched to our tests and items. The type of information that tests or groups of items provided was related to the reliability of the test (i.e., the amount of error present in the measurement of students' ability; Embretson & Reise, 2000). Item information function graphs illustrate how much information an item or groups of items provide and for what range of student abilities (Embretson & Reise, 2000). In item information function graphs, the *x*-axis represents student ability (theta) and the *y*-axis represents the amount of information provided. Thus a peak centered at theta = 0 (with 0 = average ability) would represent that this group of items provides the most information for students of an average ability level. The more information a test provides for a certain student ability (theta), the more sensitive that group of items is in determining that student's proficiency. Often tests are designed to provide the most information about students with average ability (theta) and have less information about students who are either above or below average. However, we designed our scaffold-rich assessments specifically to gather evidence of students' thinking at a range of knowledge and ability levels. Thus we hoped these assessments would provide information at various difficulty levels and therefore enhance our ability to reliably assess how well students with a range of abilities were able to construct scientific explanations.

In our assessment study, we examined the test information function graphs for our fourth, fifth, and sixth grade tests (and the different levels of scaffolded items in each test) at pre- and posttest time points for each grade level. While the pre- and posttest for each grade were identical, we conducted psychometric analysis of the pretest and posttest separately to determine how our items functioned before and after curricular implementation. Figures 5–7 present the item information function graphs for our fourth, fifth, and sixth grade tests.

Our fourth grade results demonstrated that the test information function graphs at the pretest (Figure 5a) and posttest (Figure 5b) are both centered close to theta, indicating that at both the pre- and posttest time points, the test provided the most information about students who are approximately average ability (close to theta = 0). In addition, the shape and size of the test information function graphs for fourth grade pre- and posttest looked very similar to each other. This result suggested that our pre- and posttest provide similar amounts of information on fourth graders' abilities both before and after our intervention. Concerning assessment item types, the intermediate II items, which have only explanation construction scaffolds but not content scaffolds, provided the most information followed by the complex items with no scaffolding at all. Minimal items provided significantly less information than the other three types of scaffold items, suggesting that this type of scaffolding was not an optimal means to elicit information from a range of fourth graders.

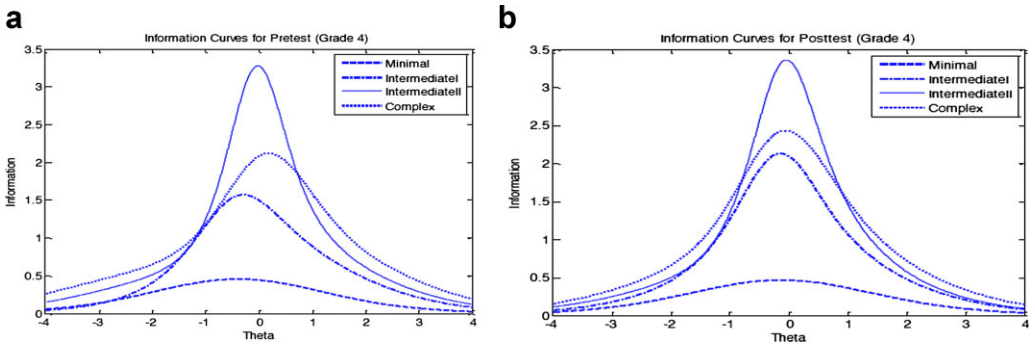

Figure 5. Test information curves for the fourth grade test at the pre- (a) and post- (b) time points.

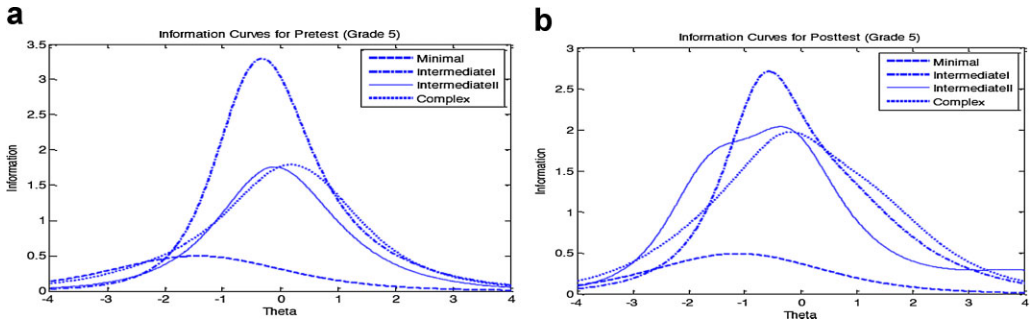

Figure 6. Test information curves for the fifth grade test at the pre- (a) and post- (b) time points.

Figure 6 presents the item information function graphs for the fifth grade pretest (Figure 6a) and fifth grade posttest (Figure 6b). Observations of the fifth grade graphs provided evidence that that the fifth grade test is a better test for a wider range of abilities than the fourth grade test (evident by the wider range of information across theta values). In particular, the fifth grade scaffold-rich items, including the intermediate I and II items with scaffolds and minimal items provided more information across a range of thetas than the fourth grade tests. A second important observation is that the fifth grade posttest provided more information than the fifth grade pretest. As both tests are identical, this result suggests that the fifth grade test is working more optimally to elicit information about fifth graders' knowledge after the intervention. Overall, the intermediate I items provided the most information across a range of theta indicating that the items that ask students to construct a full explanation with both content and explanation construction scaffolds are a good means for gathering information from fifth grade students. The items with the least scaffolding, the complex items, provided less information about fifth grade students' abilities.

Figure 7 presents the test information function graphs for the sixth grade test at the pre- (Figure 7a) and posttest (Figure 7b) time points. There are different patterns with the

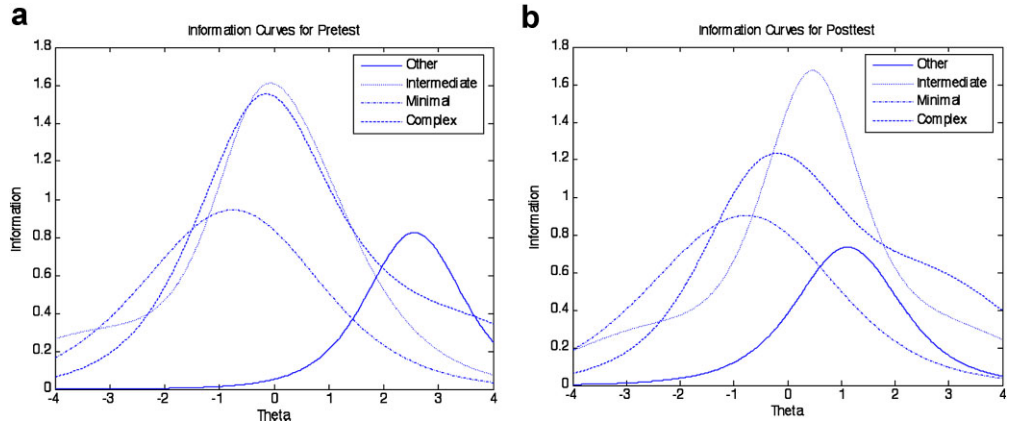

Figure 7. Test information curves for the sixth grade test at the pre- (a) and post-time points (b). Note: The scale on the *y*-axis is not as high for the sixth grade tests. The factors that influenced the shape of the test information function are complex with polytomous (partial credit) data, with issues such as the spread of category thresholds and the number of reversals in the intersection parameters. (For more information, please see Embretson and Reise, 2000; Murkaki, 1993.)

sixth grade tests than with the fourth and fifth grade tests. For example, the sixth grade tests demonstrated that different categories of items provide better information for students of different ability levels. In other words, the food web item (Other in Figure 7) provided the most information for students of higher abilities, while the minimal items provided the most information for students of below average ability. The intermediate and complex items provided the most information for students close to average ability (theta).

Another observation is the amount of information provided on posttests by the complex items. In contrast to the fifth grade posttest where the complex items provided less information on students' abilities, the complex items on the sixth grade posttest provided the most information about sixth grade students with an average ability, as they also provided the best information about students at a slightly higher ability level (note the tail that goes above the intermediate items at higher thetas). In addition, the differential information provided in the sixth grade pre- and posttest demonstrates a test that is sensitive to the curricular intervention.

The information provided by the sixth grade test therefore included information that was strong across a range of different thetas, contained categories of items that were optimal for different ability students, and that was sensitive to instruction.

The evidence that we gathered from the test information graphs indicates that all of the tests provided good information about students across a range of ability levels. However, the tests were better at providing information across a wider range of ability levels (i.e., across a wider range of thetas) with each increasing grade level. The fifth grade and sixth grade tests did a better job distinguishing students with a range of abilities to construct explanations. In both the fifth and sixth tests, the intermediate items allowed students with both below and above average ability levels opportunities to demonstrate their ability to build scientific explanations about focal content. Overall, the sixth grade test demonstrated the strongest amount of information across the widest ability levels. The minimal items provided a good amount of information for low ability students, while the intermediate and complex items provided strong information for students at average and above average levels. In this way, the sixth grade test did the best job of providing information that distinguished between students with different abilities.

Recognizing the need to gather more information about our younger audiences to help us optimize future test development, we conducted the following achievement and type of errors studies to characterize both the impact of our curricular intervention with the three audiences and to study the types of errors that were common among our youngest audience, fourth and fifth graders.

Results: Achievement and Type of Errors Studies

The second set of research studies investigated the learning outcomes and types of errors demonstrated by late elementary students in their first systematic experience with scaffold-rich explanation construction about focal concepts in biodiversity and ecology. We conducted three analyses in this section to examine achievement effects in all three grades and characterize the types of errors students demonstrated in their first attempts to construct scientific explanations in fourth and fifth grades. Multiple regression analysis was used to examine the impact of the curricular intervention (treatment) on achievement in each grade level.[1] Wright Maps, produced with the Construct Map software (Kennedy, Wilson, Draney, Tutunciyan, & Vorp, 2008) were used to examine patterns in student achievement relative to the complexity and difficulty of the items. In recognition that our fourth grade test was not an optimal assessment for eliciting information about explanation construction, we gathered information on a smaller sample of fourth and fifth grade students and conducted a Type of Errors Study in

order to examine the nature and types of errors young students exhibited in the development of explanations under scaffold conditions.

Regression Results

Table 5 presents regression results for fourth, fifth, and sixth grade achievement relative to the treatment variable (amount of the overall program completed). Table 6 presents the regression results for fourth, fifth, and sixth grade achievement relative to the amount of scaffold-rich explanation activities completed.[2] As the scaffold-rich explanation activities were a subset of the total activities (and thus completion rates are highly correlated), we ran two separate regressions for each grade level examining the impact of completion of all types of curricular activities completed (Table 5) and the number of scaffold explanation activities completed (Table 6) on student learning.

In all grade levels, students' completion of both the total number of activities and the number of scaffold-rich explanation activities were significant predictors of their learning. The results in Table 6 illustrate that for all grades, the effect size was stronger for the scaffold-rich explanation activities completion than the overall program completion. This result suggests that scaffold explanation activities were associated with stronger learning outcomes than other aspects of our units. Taking into account the amount of the curricular intervention was an important consideration in the interpretation of these results. For example, it was interesting that our youngest students (i.e., in fourth and fifth grade) were able to make significant progress from pre to posttest in demonstrating their ability to construct scientific explanations about focal content even in the cases where they had experienced less than a complete amount of the scaffold explanation activities.

Table 5
Results of fourth (N = 455), fifth (N = 294), and sixth (N = 190) grade achievement relative to program completion with posttest as the dependent variable

	4th Grade Effect Size	5th Grade Effect Size	6th Grade Effect Size
Pretest	0.178^{***}	0.139^{***}	0.456^{***}
Worksheets	0.055^{***}	0.034^{***}	$0.109^{\sim}$
R^2 value	0.493	0.45	0.254

$^{\sim}p < 0.1$.
$^{***}p \leq 0.001$.

Table 6
Results of fourth (N = 455), fifth (N = 294), and sixth (N = 190) grade achievement relative to the amount of scaffold worksheets completed with posttest as the dependent variable

	4th Grade Effect Size	5th Grade Effect Size	6th Grade Effect Size
Pretest	0.178^{***}	0.139^{***}	0.448^{***}
Explanation worksheets	0.128^{***}	0.093^{***}	0.184^{*}
R^2 value	0.493	0.44	0.276

$^{*}p \leq 0.05$.
$^{***}p \leq 0.001$.

Item Difficulty Information-Wright Map Results

As defined by the BEAR group at the University of California, Berkeley, a Wright Map is ''an aggregate map of all students' current proficiency levels versus all of the item difficulties, oriented on the same logit scale'' (BEAR, 2006). Wright Maps are a valuable analytical tool because they can provide representational information simultaneously on both the difficulty of assessment items and the performance of individuals on those same items. We used a Wright Map analyses to characterize both how difficult our assessments might be for our target audience, and to illustrate the manner in which student performance shifts from pre- to posttest.

To conduct the analysis for the generation of our Wright Maps, we calibrated the difficulty levels of all our items using our baseline data (Table 2) using a partial credit Rasch model to estimate the difficulty level of the items. The item difficulty parameters were anchored, followed by a run of the fourth, fifth, or sixth grade pre- and post-test data to obtain student ability levels (theta) relative to the item difficulty. Figures 8–10 present Wright Maps for our fourth, fifth, and sixth grade tests at the pre- and posttime points.

The Wright Maps for each grade's pre- and posttest illustrate some interesting trends. First, the Wright Maps for the fourth grade pretest (Figure 8) illustrate that only twelve students (each X represents four students) are at or above average ability level (theta; 0 = average). However, by the posttest time point, there was a movement of students to above average thetas. This movement of students upward in ability level illustrates learning correlated with the curricular intervention. However, the upward movement is slight and many students still fall below average ability, perhaps suggesting that many of the fourth graders had not experienced enough of the scaffold activities to guide them appropriately towards strong achievement on explanation construction by the posttest time point. Our data on the completion percentages of the fourth grade program show relatively lower completion rates (average of 38%), a trend we attribute to this first intervention year with fourth grade classrooms and the very small percentage of time allocated to science instruction in schools influenced strongly by No Child Left Behind pressures to prioritize reading and mathematics (NCLB, 2002). In addition, because nine items were completely out of reach on the fourth grade pretest and many of the students still fell below average ability levels at the posttest, this result also indicated that the assessment items might still be too difficult for fourth grade students even after they complete the curriculum. Our research continues to investigate in what ways we might optimize our fourth grade test to be more sensitive to a range of fourth grade ability levels.

Wright Maps for fifth grade (Figure 9) and sixth grade (Figure 10) demonstrate more obvious shifts in students' improvements from pre- to posttest time points. At the pretest time point in both fifth and sixth grade, the largest number of students cluster below a theta of 0 in both fifth grade (theta = −0.3) and sixth grade (theta = −0.2). However, by posttest time points, both populations demonstrate large numbers of students above theta = 0. These shifts in the numbers of students who were successful on the more difficult items suggest greater ability to construct explanations about focal content after the curricular intervention.

Types of Errors Results

Drawing from our Assessment Study and Wright Map analyses that indicated that our fourth grade test, in particular, was too difficult for students, we designed a study that could gather information to help us characterize the nature of errors that were common in fourth and fifth grade students' first attempts at explanation construction. For this study, we

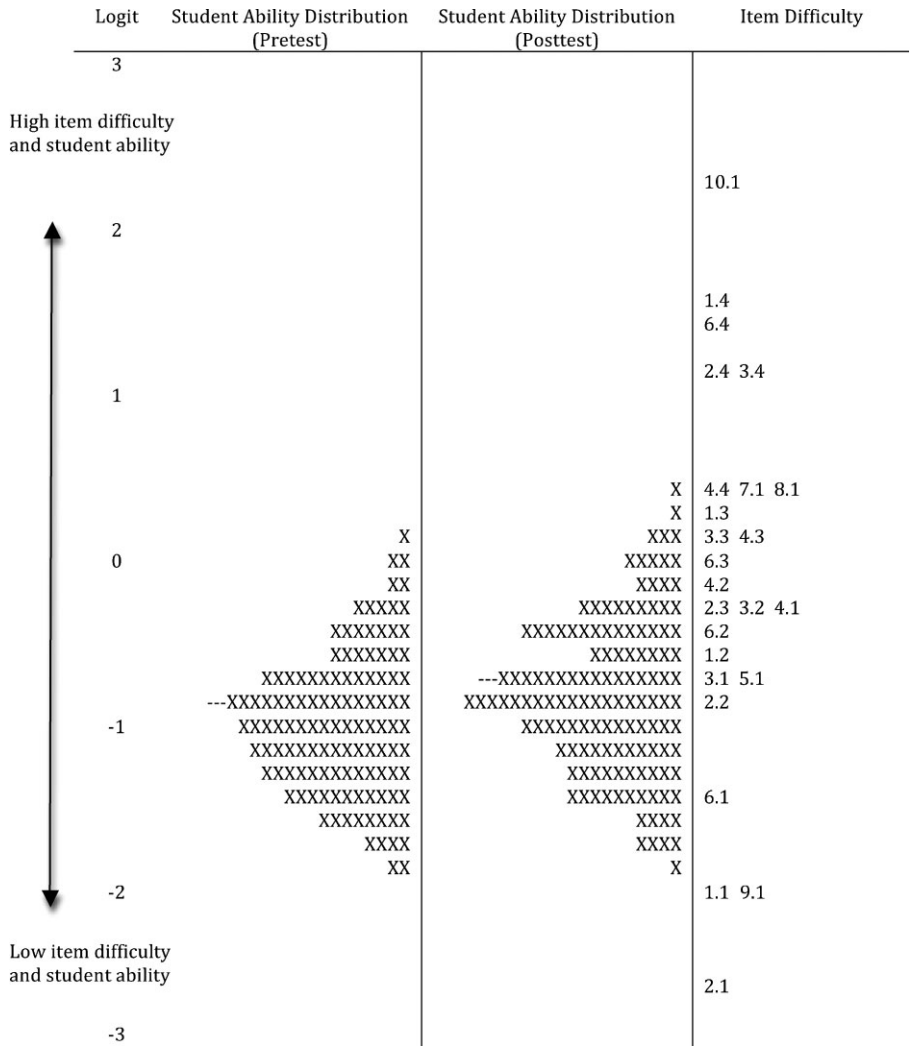

Figure 8. Fourth grade Wright Maps, with fourth pretest (a) and fourth posttest (b). Notes: Each X represents four students; each row is 0.128 logits. Item difficult is represented by Thurstonian Thresholds. The Thurstonian threshold for a particular step on an item is the ability level at which respondents have an equal chance of achieving below that step or at or above that step (BEAR).

randomly selected 27 fourth grade notebooks and 50 fifth grade notebooks from classes where students had completed at least 50% of the treatment (curricular intervention). Using a coding rubric distinguishing the type of errors observed, we examined three attempts at explanation construction in each student notebook, one from early in the intervention, one at the midpoint, and one towards the end of the curricular unit. The early intervention attempt focused on students' ability to generate a claim and evidence to match a given scientific question (worksheet 9 in fourth grade and worksheet 10 and 15 in fifth grade). The mid and later intervention attempt focused on students' ability to generate a full explanation, with a valid claim, two pieces of evidence, and reasoning that were matched to the given scientific

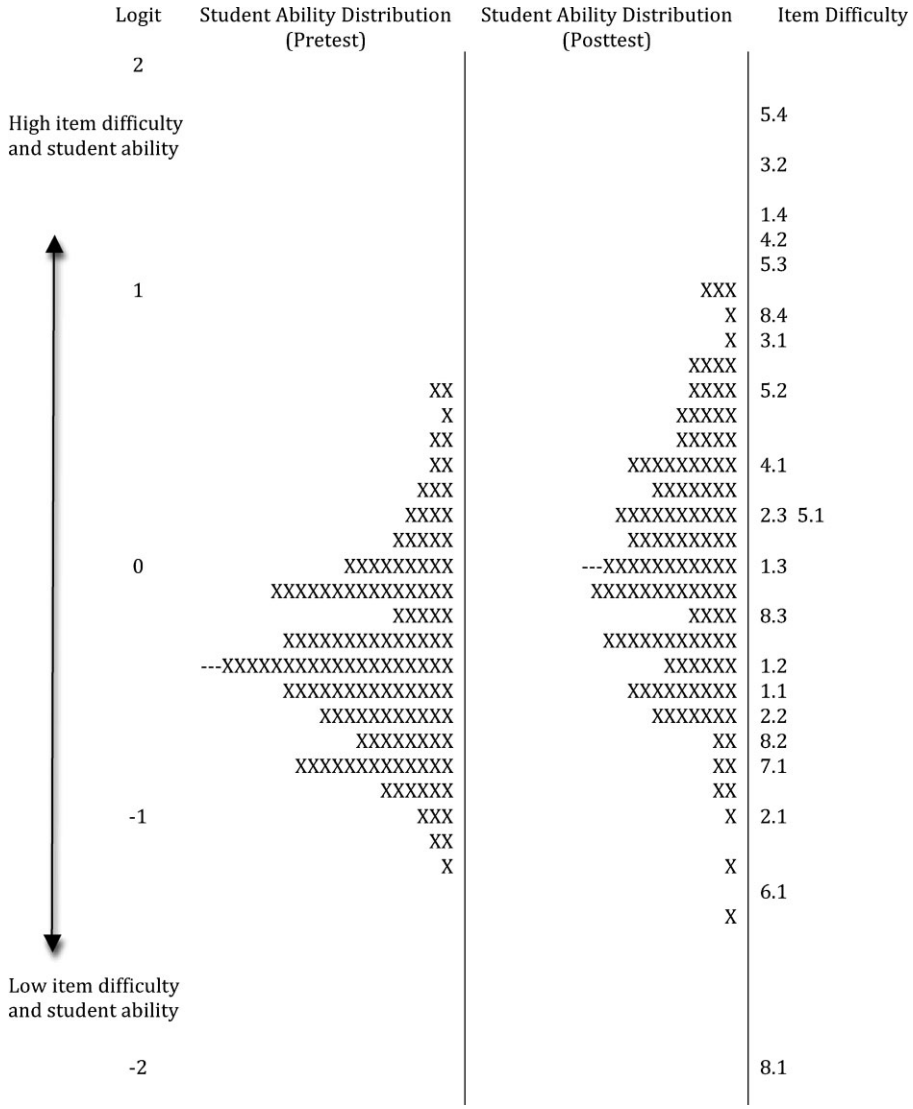

Figure 9. Fifth grade Wright Maps, with pretest (a) and posttest (b). Notes: Each X represents two students; each row is 0.096 logits. Item difficult is represented by Thurstonian Thresholds. The Thurstonian threshold for a particular step on an item is the ability level at which respondents have an equal chance of achieving below that step or at or above that step (BEAR).

question (represented as worksheet 18 and 20 in fourth grade and worksheet 29 in fifth grade). Table 7 presents our results. Note that explanation worksheets contain fewer scaffolds as the units developed, providing some explanation as to why some of the error rates were higher in activities at a later time point in the curricular unit.

As illustrated in Table 7, while there was a great deal of variety in students' responses on the scaffold explanation activities, fifth grade students in general had more fully correct explanations than fourth graders. Fourth grade students also left more answers blank than fifth

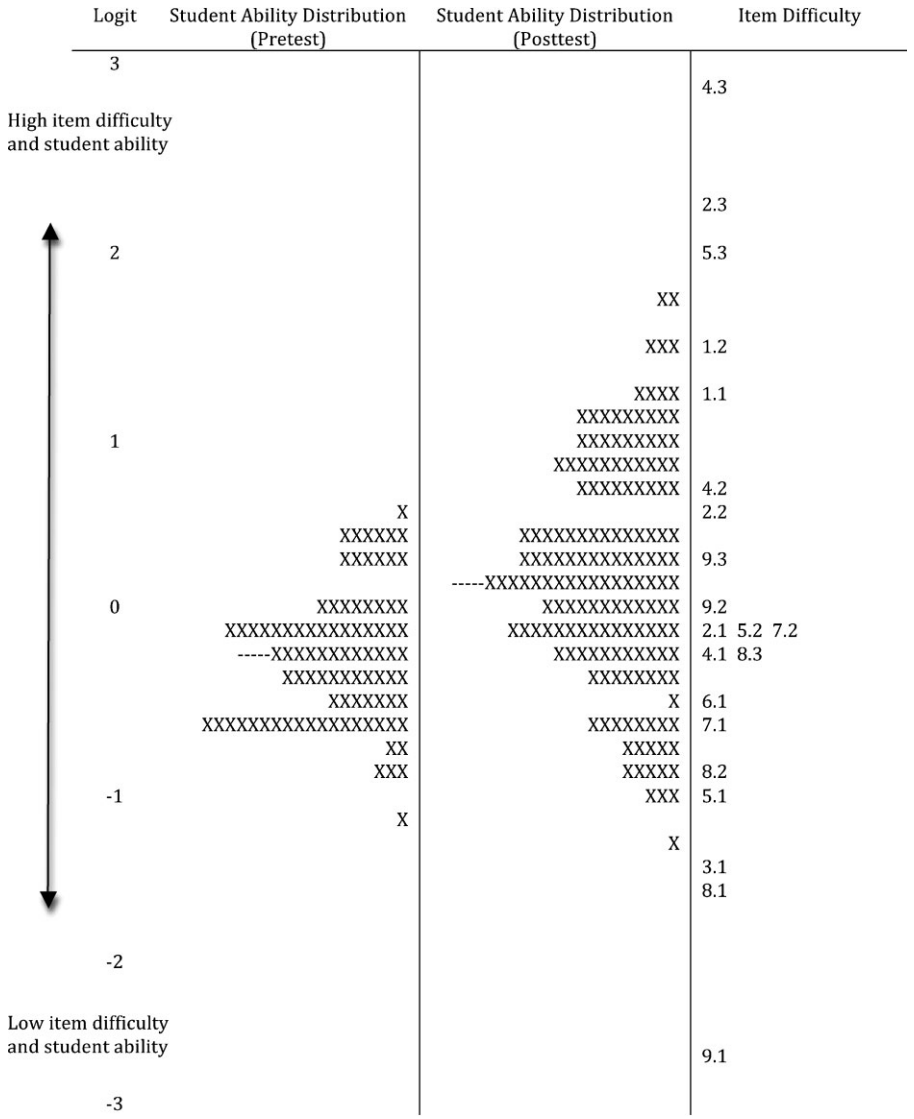

Figure 10. Sixth grade Wright Maps, with pretest (a) and posttest (b). Notes: Each X represents two students (pre) and one student (post); each row is 0.128 logits. Item difficult is represented by Thurstonian Thresholds. The Thurstonian threshold for a particular step on an item is the ability level at which respondents have an equal chance of achieving below that step or at or above that step (BEAR).

graders, even with scaffolds provided in all cases. We speculate that this type of error could be explained in a variety of ways. Our explanation activities contain a large amount of reading and required a great deal of writing. We expect that both the amount and difficulty of the vocabulary and writing tasks were overwhelming for some fourth grade students. In addition, many fourth grade students were likely overwhelmed with the new content and therefore may not have felt confident providing guesses to the claim, evidence, and reasoning prompts, even under guided conditions. While fourth graders made a great deal more errors in claims than

Table 7
Fourth and fifth grade results type of explanation error study

Grade/Time/ (Activity)	*N*	Fully Correct Explanation (%)	No Claim (%)	No Evidence (%)	No Reasoning (%)	Wrong Location (%)	Claim Present But Wrong (%)	Evidence Present But Wrong (%)	Reasoning Present But Wrong (%)
4th/early (9)	27	11.1	3.7	3.7	NA	0	11.1	44.4	NA
4th/mid (18)	27	11.1	0	33.3	14.8	7.4	25.9	33.3	29.6
4th/late (20)	8	12.5	12.5	12.5	12.5	25	37.5	37.5	50
5th/early (10)	50	76	0	0	NA	0	0	2	NA
5th/mid (15)	50	64	4	4	NA	6	0	2	NA
5th/late (29)	50	10	0	0	6	66	2	10	70

Fourth grade activity 9 (scaffold claim and evidence only); activity 18 and 20 (scaffold explanation building: claim, evidence, and reasoning). Fifth grade activity 10 and 15 (claim and evidence only); activity 29 (scaffold explanation building: claim, evidence, and reasoning).

fifth graders, both populations made the fewest amount of errors in the generation of scaffold-supported claims at all three time points. Figure 11 (left) and Figure 3 provide examples of fourth grade scaffold-rich claim and evidence activities. These fourth grade claim constructions had a high amount of supports, with students being asked only to add a number or an animal group to complete the claim sentence. Despite this high level of scaffolding, 25.9% of fourth graders still got the claim in Figure 11 (left) incorrect at the mid time point. Note that generating a portion of a claim is generally considered less difficult than constructing the entire claim, even with guidance, as is pictured in Figure 11 (right).

The data also illustrate that the parts of explanation construction that were most difficult differed for fourth and fifth graders. The most difficult aspect of explanation construction for

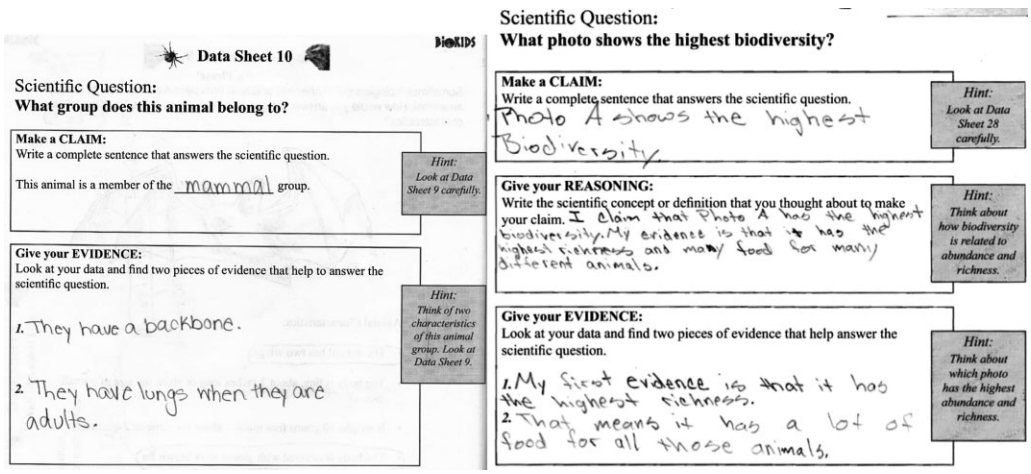

Figure 11. Left: Sample answers from week 2 activity with scaffold claim and evidence and only one correct piece of evidence (the first one). Right: Sample answer from week 7 illustrating both wrong location error (e.g., claim and evidence in the reasoning box) and only one piece of evidence (the second piece of evidence merely defines the first piece).

fourth graders was providing two pieces of valid evidence that matched their claims. Fifth graders had the most difficulty in generating reasoning that linked claims to evidence. Figure 11 illustrates examples of both these trends. The example on the left illustrates a common error where a fourth grade student only included one correct piece of evidence, despite generating two responses. (The second example is incorrect because having lungs as adults is not a characteristic of mammals.) The right side example from fifth grade illustrates two types of errors. The first error is that the student put both claim and evidence in the reasoning box, but no reasoning in the reasoning box (location errors). This outcome suggested that the student either does not understand what reasoning is, or does not know the difference between evidence and reasoning. Previous research has confirmed these as common errors in early attempts of sixth graders' explanation construction (Gotwals et al., in press). The second error was that the student only had one piece of correct evidence, as the second piece of evidence was a definition of the first piece of evidence. Careful analysis of these kinds of examples helps guide us towards more effective scaffolds and plans for fading scaffolds in future versions of our curricular activities.

Discussion and Implications

This article began with a recognition that policy documents and science standards of the 1990s presented exciting but idealized illustrations of science inquiry, including elementary students' construction of explanations about focal science content. On balance, the presentation of separate lists of content and practice (inquiry) standards with no guidelines on how to fuse explanation building with content left some confusion and ambiguity on prescriptions for guiding students, particularly elementary-age students, in explanation knowledge development. With the introduction of approaches, such as learning progressions, and new ideas in science assessment, such as scaffold-rich assessments, science educators are better positioned to investigate students' learning over time. However, it is essential to evaluate the strengths and weaknesses of learning progression-based resources such as consecutive curricular units and assessment instruments that can guide and evaluate students' explanation construction of focal science, even at the entry points of the learning progression.

Assessment Design and Implications

Our studies with assessment design were framed by a desire to develop instruments to gather empirical data about three cohorts of young students' abilities to fuse focal content with explanation building. Our goal was to design three coordinated assessment instruments, each of which provided a great deal of information on a range of different students with different knowledge and ability levels within the target grade level audience of fourth, fifth, or sixth graders. Our results from our assessment study illustrate test profiles suggesting that, even as we see room for improvement, our tests provide information on a range of late elementary students' ability to fuse content with explanation building with various amounts and types of cognitive scaffolds. Our results provided evidence that our sixth grade test was the most ideal because it not only contained categories of scaffold-rich items that provided strong information for students with a range of different ability levels, but it demonstrated sensitivity to the treatment intervention. In relation to our work, we believe that tests that demonstrate curricular sensitivity are more valuable assessment instruments than those that do not because tests with curricular sensitivity can serve as an valid evaluation of the impact of the curricular intervention (National Research Council, 2004).

On balance, as our fourth grade test did not provide as much information on a range of fourth grader's abilities nor was it as sensitive to our curricular intervention, it was not seen

as a strong test. While our results from the type of error study provide important guidance, more information is needed to decisively determine if our fourth grade tests were less strong because of the lower amounts of curricular implementation among fourth grade teachers or because too many test items were too difficult for a majority of our target fourth grade audience.

Several implications for test improvement and design are evident from our results. Perhaps, in creating explanation assessment for younger students, it may be more helpful to have questions that do not force students into the way that we view scientific explanations. Rather, a more open ended assessment that places less emphasis on abstract definitions of claim, evidence and reasoning or that uses a teacher's specific redefinitions of evidence, may prove to be helpful in implementing similar explanation construction activities in the future. Such an approach might help us to learn more about how students appropriated the scientific explanation framework that was presented in the curriculum and made it part of their own framework for explaining. In other words, with a prompt for evidence such as, ''how would you support this claim?'' we could provide students with less abstract support such as the verbal scaffolds provided by our teachers (Songer, Shah and Fick, in press).

Achievement, Errors and Implications

In our curricula activity development and evaluation, we struggled to gather empirical data to address questions such as: What does it mean to break down scientific explanations into smaller pieces for more successful student explanation construction? When we do attempt to simplify and provide guidance in student construction of such a complicated science practice, what does success and failure look like, and how do we build on these outcomes towards more fruitful outcomes? Our achievement results demonstrated strong results across all three grades when taking into account the amount of the curricular intervention and the amount of the scaffold-rich activities each class implemented. An interesting aspect of this outcome was that our achievement results demonstrated that our younger students were able to make significant progress from pre to posttest in demonstrating their ability to fuse content knowledge with explanation construction even in the cases where they had experienced less than a complete amount of the scaffold explanation activities. These results are important, particularly in relation to our Wright Map results that demonstrated less curricular sensitivity of our fourth grade test relative to our fifth and sixth grade tests.

Type of error studies demonstrated higher percentages of fourth graders leaving claim, evidence and reasoning boxes blank on their activity sheets. A likely explanation for this result is the large amounts of reading and writing that may impede our fourth graders from taking maximum advantage of the cognitive supports provided. In addition, fourth graders demonstrated the most difficulty in generating evidence. Research by McNeill (2011) suggests another possible explanation for our fourth graders' weakness in generating two pieces of evidence that supports their claim: that fourth graders do not recognize evidence as data.

Our results suggest several implications for how to improve our scaffold-rich activities for elementary students. First, our results suggest a reconsideration of the amount of reading and writing presented to younger audiences. Second, our results suggest slowing down the pace at which the scaffolds are faded, particularly in the fourth grade unit. Finally, our results and that of others suggested careful attention to the type of talk and teacher support that the teacher provides in conjunction with the written guides for ideas such as evidence. A recent research study designed to characterize fourth, fifth, and sixth grade teachers' talk relative to their guidance of students' explanation construction revealed interesting insights. In this study, fourth and fifth grade teachers utilized up to five different verbal variations to define

evidence throughout the eight week unit, and these definitions were revisited multiple times throughout the unit. An example of a redefinition was defining evidence as ''the data that goes with what we're looking for'' and ''proof'' (Songer et al., in press). Similarly, recent work by Lehrer and Schauble (2010) analyzed classroom discussions and the manner in which an elementary teacher reached consensus with her students about complex constructs such as ''good questions.'' In these discussions, teachers not only redefined key terms multiple times, but they referenced these definitions several times as standards of classroom dialogue.

Conclusion

We encourage more studies that can provide empirical information and guidance on how to move beyond existing standardized and off-the-shelf tests towards assessments that can help us understand what makes core science fused with explanation construction and other science practices difficult for young audiences. In addition, we encourage additional research studies that provide empirical information to guide us towards an understanding of what it means to guide younger students, in both written scaffolds and verbal teacher scaffolds, in fusing core science with science practices (Lehrer & Schauble, 2010; Songer, Shan, & Fick, in press). Through research studies that optimize content-practices assessments and that guide us towards a productive break down of critical thinking, such as explanation construction, into fruitful scaffolds, variations, and practices, we can build productively on one another's insights towards successful science knowledge development by younger students.

Notes

[1]Note that we used the imputed data for analyses.

[2]In each grade, pretest is a significant predictor of posttest ($p < 0.001$). Controlling for the pretest allows us to examine student learning gains from pretest to posttest.

References

BEAR. (2006). Berkeley Center for Evaluation and Assessment. Downloaded from http://bearcenter.berkeley.edu/kennedy/GMOnline/Wright_Maps.html on 3/16/2011.

Bond, T. G., & Fox, C. M. (2001). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: Lawrence Erlbaum Associates.

The College Board. (2009). Science: College Board Standards for College Success. New York, NY: The College Board.

Davis, E., & Krajcik, J. (2006). Designing educative curriculum materials to promote teacher learning. Educational Researcher, 34(3), 3–14.

No Child Left Behind. (2002). Public Law 107-110. 107th Congress. January 8, 2002.

Dewey, T. A., Hammond, G. S., Espinosa, R., Parr, C. S., Jones T., & Myers, P. (2011). BioKIDS Critter Catalog (online). http://www.biokids.umich.edu.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.

Gotwals, A. (2006). Students' science knowledge bases: Using assessment to paint a picture (unpublished doctoral dissertation). University of Michigan, Ann Arbor.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. Science Education, 94, 259–281.

Gotwals, A., Songer, N. B., Bullard, L. (in press). Assessing students' progressing abilities to Construct Scientific Explanations. In A. C. Alonzo & A. W. Gotwals (Eds.), Learning progressions in science: Current challenges and future directions. Rotterdam, The Netherlands: Sense Publishing.

Kennedy, C. A., Wilson, M., Draney, K., Tutunciyan, S., & Vorp, R. (2008). ConstructMap Version 4.4.0. (computer program). UC Berkeley, CA: BEAR Center.

Lee, H. S., & Songer, N. B. (2003). Making authentic science accessible to students. International Journal of Science Education, 25(1), 1–26.

Lehrer, R., & Schauble, L. (2010). What kind of explanation is a model? In M. K. Stein & L. Kucan (Eds.), Instructional explanations in the disciplines (pp. 9–22.) New York, NY: Springer.

Linn, M. C., Shear, L., Bell, P., & Slotta, J. (1999). Organizing principles for science education partnerships: Case studies of students' learning about 'rats in space' and 'deformed frogs'. Educational Technology Research and Development, 47(2), 61–84.

Linn, M. C., Bell, P., & Davis, E. (2004). Specific design principles: Elaborating the scaffold knowledge integration framework. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), Internet environments for science education (pp. 315–339). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometricka, 47, 149–174.

McNeill, K., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. Lovett & P. Shah (Eds.), Thinking with data (pp. 233–265). New York: Taylor & Francis.

McNeill, K. (2011). Elementary students' views of explanation, argumentation, and evidence and their abilities to construct arguments over the school year. Journal of Research in Science Teaching, 48(7), 793–823.

Metz, K. E. (1991). Development of explanation: Incremental and fundamental change in children's physics knowledge. Journal of Research in Science Teaching, 28(9), 785–797.

Murkaki, E. (1993). Information functions of the generalized partial credit model. Applied Psychological Measurement, 17, 351–363.

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

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

National Research Council. (2004). On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations. Washington, DC: National Academies Press.

National Research Council. (2007). Taking science to school: Learning and teaching science in grades K-8. Washington, DC: National Academies Press.

National Research Council. (2011). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: The National Academies Press.

OECD. (2007). PISA 2006: Science competencies for tomorrow's world volume 1: Analysis. Paris, France: Organisation for Economic Co-operation and Development.

Parr, C. S., Espinosa, R., Jones, T., McDonald, S., Songer, N. B., & Myers, P. (2003). Introductory-level Cyber Tracker sequence for Detroit-area wildlife, augmented by web-based data summary and display. The University of Michigan.

Partnership for 21st Century Skills. (2009). Framework for 21st Century Learning. Downloaded from www.p21.org/documents/P21_Framework.pdf on 3/16/2011.

Peters, V. Songer, N. B. (forthcoming). The co-design of interdisciplinary knowledge in science education.

Quintana, C., Reiser, B., Davis, E., Krajcik, J., Fretz, E., Duncan, R., ... Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. The Journal of the Learning Sciences, 13(3), 337–386.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 27(1), 85–96.

Reiser, B. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. The Journal of the Learning Sciences, 13(3), 273–304.

Ruberg, S. J. (1989). Contrasts for identifying the minimum effective dose. Journal of the American Statistical Association, 84(407), 816–822.

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. Journal of the Learning Sciences, 12(1), 5–51.

Songer, N. B. (2006). BioKIDS: An animated conversation on the development of curricular activity structures for inquiry science. In R. Keith Sawyer (Ed.), Cambridge Handbook of the Learning Sciences (pp. 355–369). New York: Cambridge.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). When and how does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. Journal of Research in Science Teaching, 46(6), 610–631.

Songer, N. B., Shah, A. M., Fick, S. (in press). Characterizing teachers' verbal scaffolds to guide elementary students' creation of scientific explanations. School Science and Mathematics.

Toulmin, S. (2006). The uses of argument (updated edition). New York: Cambridge University Press.

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