How and When Does Complex Reasoning Occur? Empirically Driven Development of a Learning Progression Focused on Complex Reasoning about Biodiversity

Nancy Butler Songer,[1] Ben Kelcey,[2] Amelia Wenk Gotwals[3]

[1]School of Education, The University of Michigan, Ann Arbor, Michigan
[2]College of Education, Wayne State University, Detroit, Michigan
[3]School of Education, Michigan State University, East Lansing, Michigan

Received 16 June 2008; Accepted 14 April 2009

Abstract: In order to compete in a global economy, students are going to need resources and curricula focusing on critical thinking and reasoning in science. Despite awareness for the need for complex reasoning, American students perform poorly relative to peers on international standardized tests measuring complex thinking in science. Research focusing on learning progressions is one effort to provide more coherent science curricular sequences and assessments that can be focused on complex thinking about focal science topics. This article describes an empirically driven, five-step process to develop a 3-year learning progression focusing on complex thinking about biodiversity. Our efforts resulted in empirical results and work products including: (1) a revised definition of learning progressions, (2) empirically driven, 3-year progressions for complex thinking about biodiversity, (3) an application of statistical approaches for the analysis of learning progression products, (4) Hierarchical Linear Modeling results demonstrating significant student achievement on complex thinking about biodiversity, and (5) Growth Model results demonstrating strengths and weaknesses of the first version of our curricular units. The empirical studies present information to inform both curriculum and assessment development. For curriculum development, the role of learning progressions as templates for the development of organized sequences of curricular units focused on complex science is discussed. For assessment development, learning progression-guided assessments provide a greater range and amount of information that can more reliably discriminate between students of differing abilities than a contrasting standardized assessment measure that was also focused on biodiversity content. © 2009 Wiley Periodicals, Inc. J Res Sci Teach 46: 610–631, 2009

Keywords: learning progressions; complex thinking; biology

Consistently, international test results and policy documents in the United States conclude that the American science education system is doing a poor job preparing students to be globally competitive in science and mathematics. Policy documents such as Rising Above the Gathering Storm (National Academy of Science, 2007) testify that American science and mathematics students consistently demonstrate low test scores, and the science curricula and teacher preparation programs in K-12 science education are weak as compared to other industrialized countries. The Programme for International Student Assessment (PISA; OECD, 2007) is an assessment project designed to provide policy-oriented international indicators of complex learning and applied knowledge of 15-year-old students worldwide. The 2006 PISA focus on science literacy included an evaluation of students' ability to interpret data, critique scientific evidence, and apply knowledge of scientific concepts to current topics such as DNA fingerprinting and biodiversity. On the 2006 test, American 15-year olds performed poorly overall including a rank of 29th out of 57 countries that was significantly below the OECD average. In comparison, the Canadian average ranked third overall behind Finland (first) and Hong Kong (second).

Contract grant sponsor: National Science Foundation; Contract grant numbers: REC-0089283, REC-0129331.
Correspondence to: Nancy Butler Songer; E-mail: songer@umich.edu
DOI 10.1002/tea.20313
Published online 7 July 2009 in Wiley InterScience (www.interscience.wiley.com).

© 2009 Wiley Periodicals, Inc.

One possible solution to existing problems in American science education emphasizes the idea of systematic curricular programs that build understandings of science content through organized, guided, and repeated exposures to concepts and reasoning skills over multiple curricular units and years. Aspects of this perspective are not new to science education; science educators in the 1960s such as Robert Karplus (e.g., Karplus & Their, 1967) and Jerome Bruner introduced the idea of a spiral curriculum with an emphasis on an organizational plan for the systematic presentation, revisiting and building of concepts,

"That is to say, domains of knowledge are *made*, not *found* . . . A good intuitive, practical grasp of a domain at one stage of development leads to better, earlier, and deeper thinking in the next stage when the child meets challenging new problems in that domain. As a teacher, you do not wait for readiness to happen; you foster or "scaffold" it by deepening the child's powers at the stage where you find him or her now." (Bruner, 1996, pp. 119–120)

A REVISED DEFINITION OF LEARNING PROGRESSIONS

These ideas of systematically fostering readiness and making rather than finding domains of knowledge are central foundations contained in a new curriculum and assessment design approach called *learning progressions*; an approach that has emerged as a research tool to guide both the development and the evaluation of curricular programs organized to foster more sophisticated thinking about selected, essential topics over multiple curricular units and years. Placing an emphasis on the establishment of a systematic sequence of scientific ideas that build with and on one another over time, learning progressions are defined as "successively more sophisticated ways of thinking about a topic that can follow and build on one another as children learn about a topic over a broad span of time" (National Research Council, 2007; p. 217).

While at first blush it might seem that readers will share a common understanding of this definition, the articles of this special issue confirm a healthy dialogue that includes slight variations as to the exact nature of a learning progression. Is a learning progression a sequence of science topics such as "food chains before food webs" that, if taught in the sequence specified, might lead to a more robust comprehension of the science domain? We suggest that this definition of a learning progression is potentially problematic for two reasons. First, defining a learning progression as merely a sequence of science topics oversimplifies an essential dimension of the NRC definition, "successfully more sophisticated ways of thinking about a topic . . ." (National Research Council, 2007; p. 217). "Ways of thinking about a topic" recognizes the inherent presence and interconnection of content knowledge with inquiry reasoning (Catley, Lehrer, & Reiser, 2004; Songer, 2006). Therefore defining a learning progression as only content knowledge without consideration of inquiry reasoning is problematic. As research in science education suggests that content and inquiry reasoning skills develop in concert even if the specifics of the mechanisms of development and the nature of the relationship are not clear (Gotwals & Songer, 2006; NRC, 2007), a learning progression fostering "more sophisticated ways of thinking about a topic" must include both the increasingly more sophisticated sequence of content topics and the increasingly more sophisticated progression of inquiry reasoning skills, also called scientific practices (NRC, 2007), over time. We address this point in our work with the presentation of both a content progression and an inquiry reasoning progression, which together constitute a learning progression for our focal topic.

A second related and potentially problematic aspect of some definitions of a learning progression arises in the evaluation of whether or not a learning progression is successful or not. What does it mean to "evaluate a learning progression" for evidence of success? How can a content sequence or an inquiry reasoning sequence be evaluated? We suggest that neither a content sequence nor an inquiry reasoning sequence can be directly evaluated. Instead, the content and inquiry reasoning progressions serve as a resource for the generation of products, such as curricular products, which can be empirically evaluated. In our work, learning progressions are a template for the design of curricula, assessment and professional development products, which, subsequently, can be evaluated relative to student learning outcomes.

To emphasize the inclusion of both content and inquiry reasoning sequences and the evaluation of curricular activities that are manifestations of learning progressions, we suggest an expanded definition of learning progressions as follows:

> Learning progressions take 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. Learning progressions are successively more sophisticated ways of thinking about a topic that can be used as templates for the development of curricular and assessment products. Learning progressions-driven curricular and assessment products are one of several possible manifestations of a given learning progression. The learning progression can only be evaluated indirectly, through the evaluation of the curricular products, professional development modules, and assessment instruments that are constructed from the learning progressions template.

A FIVE-STEP PROCESS OF LEARNING PROGRESSION DEVELOPMENT

In this article we present and discuss our iterative, empirically driven work to develop a learning progression focused on complex reasoning about biodiversity for fourth–sixth graders. Our work was manifested in five steps: (1) the development of a preliminary content progression and a preliminary inquiry reasoning progression to serve as a template for the development of learning progression work products (curricular units and assessment instruments); (2) the development of 8 weeks of curricular activities that were a manifestation of our first content and inquiry reasoning progressions; (3) the development of assessment items matched to our first content and inquiry reasoning progressions; (4) the evaluation of learning that occurred with our curricular manifestations through our assessment instruments matched to our learning progression, and (5) the revision and expansion of our learning progression into a 3-year content progression and 3-year inquiry reasoning progression for the development of complex thinking about biodiversity. The following sections describe each of these steps in more detail, and the products and results that were realized.

Step 1: First Content and Inquiry Reasoning Progressions

Three major decisions guided our early work in the development of preliminary content and inquiry reasoning progressions. First, we recognized the necessity of selecting and prioritizing content ideas (e.g., core ideas, NRC, 2007) that would serve as the focus of our learning progressions work. Second, we recognized the value of drawing from expert scientists' multifaceted understandings of our focus topic in the determination of focal points for learning progression development. Third, we were resolute in our belief that the development of our content and inquiry reasoning progressions should be empirically based; therefore we decided that our early ideas should be tested empirically prior to the development of our 3-year progressions. Combining these decisions, step one of our design process was to engage in lengthy discussions with scientists over approximately 8 months to articulate our first version of the essential dimensions of complex thinking about biodiversity to support the construction of our first draft of content and inquiry reasoning progressions.

Current definitions of scientific literacy emphasize complex reasoning such as knowing, using and interpreting scientific explanations, and evaluating and applying evidence and arguments appropriately (National Research Council, 2007). Drawing from the policy recognition of the importance of evidence-based explanations and our previous work with our research team of scientists and educators to guide and evaluate students' development of evidence-based explanations (e.g., Lee & Songer, 2003; Songer, 2006), we chose *evidence-based explanations* as the core idea of our first inquiry reasoning progression.

For 7 years, the research project has been working closely with zoologists to transform scientific resources such as the Animal Diversity Web designed for an adult audience into resources such as Critter Catalog that support inquiry questioning, and explanation-building by fourth–sixth grade students. We call this transformation process making the resource "simply complex" as the transformation process must maintain the integrity of selected aspects of the complexity of the science topic in order to be valuable for student questioning and explanation-building, while making many aspects of the resource and the scientific information simple enough to be usable for fourth–sixth grade audiences. Continuing our work with scientists of the team, we selected *biodiversity* as the core idea for the first content progression. Biodiversity was also selected to acknowledge the important and timely nature of the topic relative to the impact of potential global climate change shifts on organisms, populations, and species, and the relative impact on agriculture, public health, and ecological balance (e.g., Peterson, 2003).

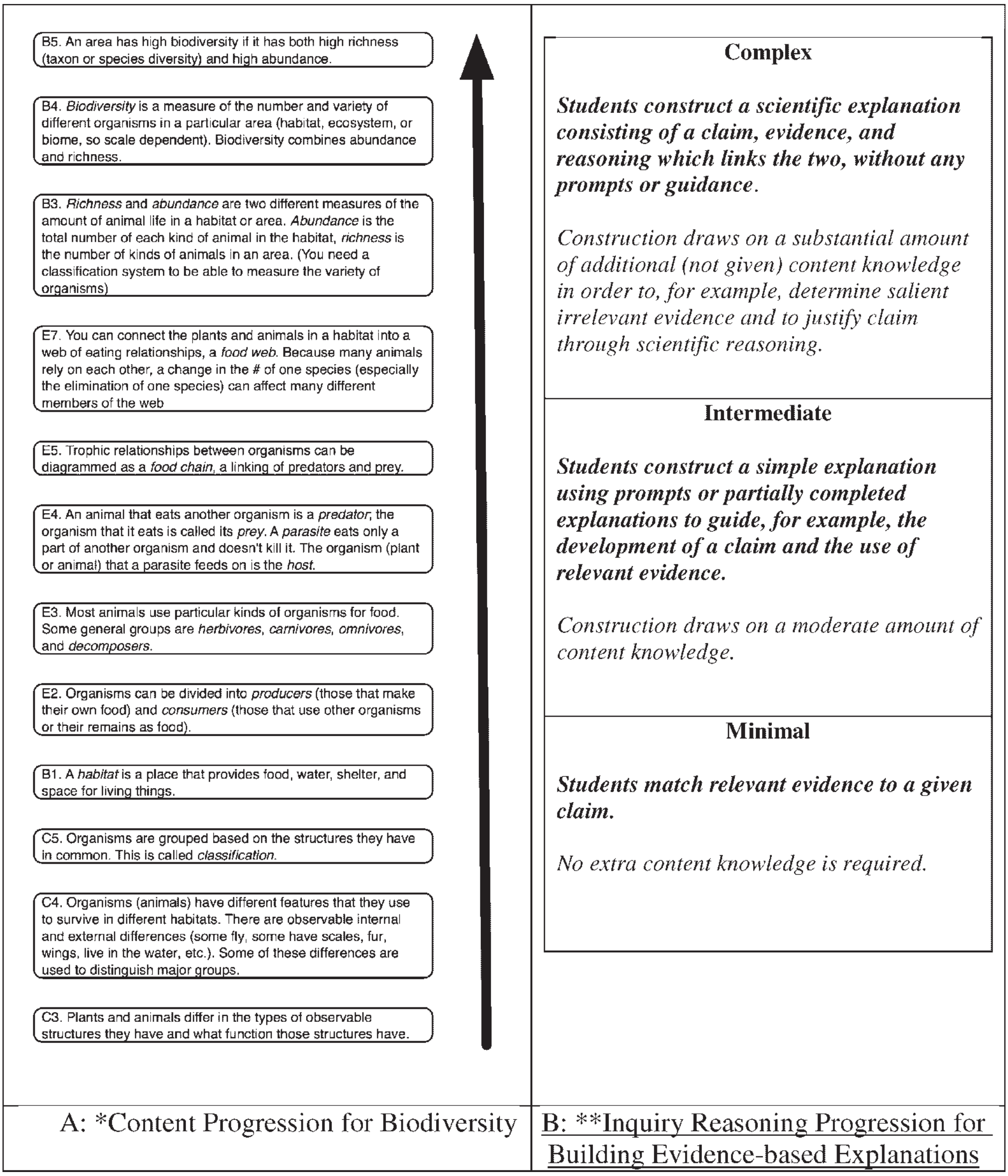
Once the selection of evidence-based explanations and biodiversity were chosen as core ideas, we engaged in extended conversations with our scientists to address national (e.g., NRC, 1996) and state science standards for sixth grade while also supporting students in developing focal aspects of biodiversity that would represent a ''simply complex'' understanding of biodiversity. Table 1 presents the first version of our content progression for biodiversity and our inquiry reasoning progression for building evidence-based explanations that arose from these conversations and that together constitute our learning progression templates associated with *building evidence-based explanations about biodiversity.* Our version one progressions contain the simpler or first ideas at the bottom and the more complex or advanced ideas at the top. The content progression contains 12 focal points in the sub areas of C: Classification, B: Biodiversity, and E: Ecology, while our first inquiry progression contains three levels: minimal, intermediate, and complex. The three levels of inquiry reasoning associated with building evidence-based explanations takes as a foundation a definition of scientific explanation that draws from the work of Toulmin (1958) and that was developed in concert with our early work in inquiry assessment in conjunction with the Principled Assessment Design for Inquiry (PADI) project led by Geneva Haertel and Robert Mislevy (Gotwals & Songer, 2006).

Recognizing the essential integration of content and inquiry reasoning knowledge for the development of complex thinking about a focal topic, our work emphasizes that every learning progression product that is developed as a manifestation of our progressions must reference a focal point on both a content progression and an inquiry reasoning progression. Central to our thinking is the working hypothesis that content and inquiry reasoning progressions exist as parallel templates that together constitute a learning progression for a focal topic, however despite this parallel presentation, inquiry reasoning and content are never considered as separate learning goals. In our work we did not integrate the content and inquiry reasoning progressions into one template to acknowledge our previous work (Songer, 2006) that suggests that the fostering of ''more sophisticated way of thinking about a topic'' might suggest a cyclical path along our inquiry reasoning progression even if it suggests a linear path along our content progression. In other words, in an ideal curricular unit manifested from our progressions, students could be working with one level of the inquiry reasoning progression (e.g., intermediate) many times in combination with different focal points along the content progression. An ideal curricular unit might guide students to do intermediate inquiry reasoning associate with C3: plant and animal structure and function followed by intermediate inquiry reasoning associated with C4: features and survival in different habitats. As we hypothesize that there is not one ideal manifestation of the content plus inquiry reasoning pairings that conforms to a linear progression, our plan was to define ideal templates for content and inquiry reasoning progressions, then develop learning progression manifestations (e.g., curricular and assessment products) to empirically test our work.

Step 2: Curricular Activities That Manifest Content and Inquiry Reasoning Progressions

Step two of our design process was to systematically translate focal points from our content and inquiry reasoning progressions into curricular activities to be implemented with students and tested empirically. Our curriculum design process drew from learning theories rooted in constructivism (e.g., Inhelder and Piaget, 1958) and the idea that fostering ''simply complex'' science was achievable through an organized developmental progression of activities that includes higher-order thinking even at younger ages (Metz, 2000). A central dimension of our work was drawing from existing work in cognitive scaffolds (Quintana et al., 2004; Reiser, 2004; Lee and Songer, 2003) and working with Detroit teachers to develop a scaffold format for students' first guided development of evidence-based explanations associated with our focal concepts. Previous work with three versions of our explanation-building format led to an understanding of the necessity of providing specific locations (such as boxes in this version) for students' inclusion of the components we defined as essential in an evidence-based explanation: a scientific claim, two pieces of evidence (associated with a key scientific concept), reasoning that ties the claim to the evidence, and guidance in composing all of these pieces into one coherent whole (Songer, 2006). Our curricular unit contained ten examples of the explanation-building format each associated with a different focal topic from the content progression. See Figure 1 for one example of a curricular worksheet that uses the explanation-building format to guide students to build an evidence-based explanation as to whether or not and why a given animal is an insect.

Table 1
First progressions for building evidence-based explanations about biodiversity including (a) content progression for biodiversity and (b) inquiry reasoning progression



B5. An area has high biodiversity if it has both high richness (taxon or species diversity) and high abundance.

B4. *Biodiversity* is a measure of the number and variety of different organisms in a particular area (habitat, ecosystem, or biome, so scale dependent). Biodiversity combines abundance and richness.

B3. *Richness* and *abundance* are two different measures of the amount of animal life in a habitat or area. *Abundance* is the total number of each kind of animal in the habitat, *richness* is the number of kinds of animals in an area. (You need a classification system to be able to measure the variety of organisms)

E7. You can connect the plants and animals in a habitat into a web of eating relationships, a *food web*. Because many animals rely on each other, a change in the # of one species (especially the elimination of one species) can affect many different members of the web

E5. Trophic relationships between organisms can be diagrammed as a *food chain*, a linking of predators and prey.

E4. An animal that eats another organism is a *predator*; the organism that it eats is called its *prey*. A *parasite* eats only a part of another organism and doesn't kill it. The organism (plant or animal) that a parasite feeds on is the *host*.

E3. Most animals use particular kinds of organisms for food. Some general groups are *herbivores*, *carnivores*, *omnivores*, and *decomposers*.

E2. Organisms can be divided into *producers* (those that make their own food) and *consumers* (those that use other organisms or their remains as food).

B1. A *habitat* is a place that provides food, water, shelter, and space for living things.

C5. Organisms are grouped based on the structures they have in common. This is called *classification*.

C4. Organisms (animals) have different features that they use to survive in different habitats. There are observable internal and external differences (some fly, some have scales, fur, wings, live in the water, etc.). Some of these differences are used to distinguish major groups.

C3. Plants and animals differ in the types of observable structures they have and what function those structures have.

Complex

Students construct a scientific explanation consisting of a claim, evidence, and reasoning which links the two, without any prompts or guidance.

Construction draws on a substantial amount of additional (not given) content knowledge in order to, for example, determine salient irrelevant evidence and to justify claim through scientific reasoning.

Intermediate

Students construct a simple explanation using prompts or partially completed explanations to guide, for example, the development of a claim and the use of relevant evidence.

Construction draws on a moderate amount of content knowledge.

Minimal

Students match relevant evidence to a given claim.

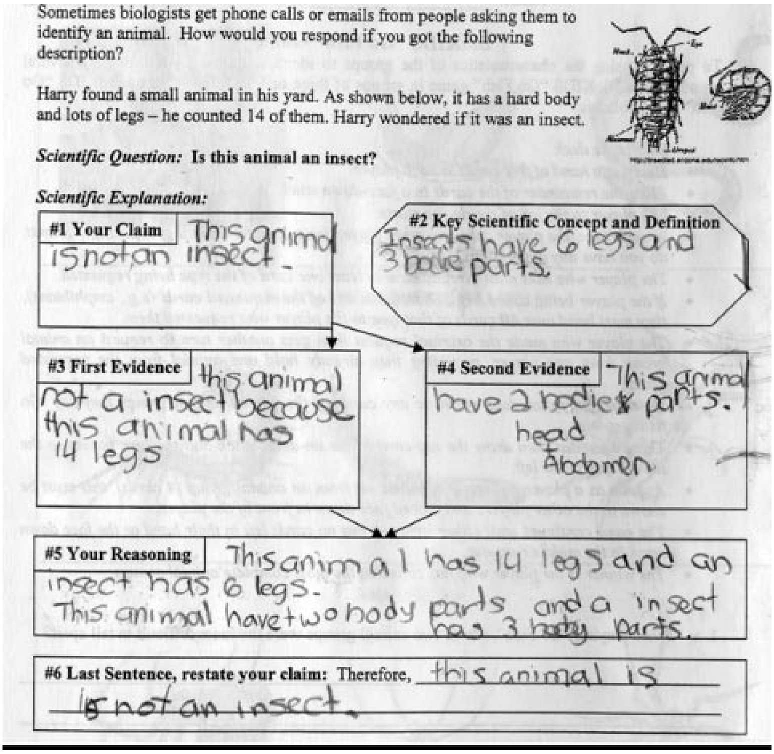
No extra content knowledge is required.

A: *Content Progression for Biodiversity

B: **Inquiry Reasoning Progression for Building Evidence-based Explanations

*C: Classification, E: Ecology and B: Biodiversity letters and numbers refer to similar reference points in Table 9A: 3-year content progression for biodiversity.

**Explanations consist of a claim, evidence and reasoning where: Claim—answer to a scientific question, Evidence—data that support the claim (Evidence should be appropriate/relevant, sufficient (enough evidence is used to support the claim)), and Reasoning—use of scientific principles to tie the claim and evidence together.



Sometimes biologists get phone calls or emails from people asking them to identify an animal. How would you respond if you got the following description?

Harry found a small animal in his yard. As shown below, it has a hard body and lots of legs – he counted 14 of them. Harry wondered if it was an insect.

Scientific Question: **Is this animal an insect?**

Scientific Explanation:

#1 Your Claim This animal is not an insect.

#2 Key Scientific Concept and Definition Insects have 6 legs and 3 bodie parts.

#3 First Evidence this animal not a insect because this animal has 14 legs

#4 Second Evidence This animal have 2 bodies parts. head Abdomen

#5 Your Reasoning This animal has 14 legs and an insect has 6 legs. This animal have two body parts and a insect has 3 body parts.

#6 Last Sentence, restate your claim: Therefore, this animal is is not an insect.

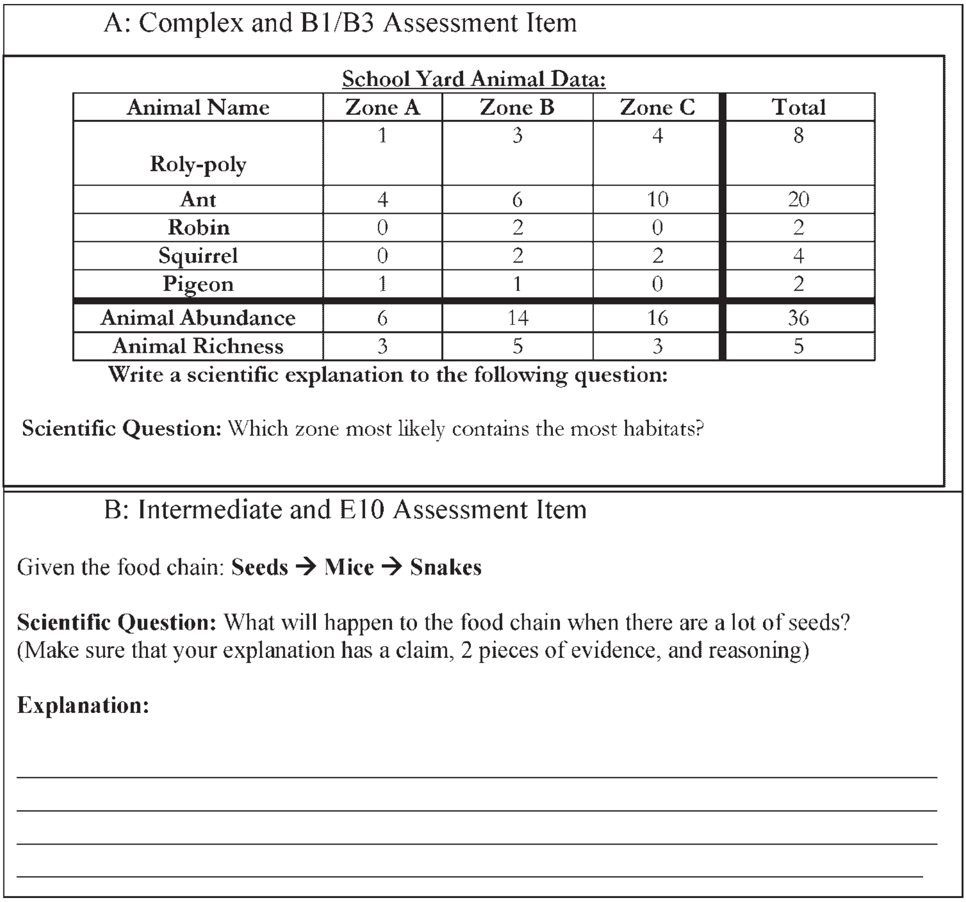
Figure 1. Explanation-building format associated with intermediate inquiry reasoning and C4 content.

Step 3: Learning Progression-Mapped Development of Assessment Items

Our assessment design followed principles of Evidence Centered Design (ECD; Mislevy, Almond, & Lukas, 2004) towards the creation of pretests, embedded assessments, and posttests calibrated on the same scale. A major tenet of the ECD approach is the recognition that assessment is a vehicle for the gathering of empirical evidence so that we can make inferences about unobserved phenomena; in this case our content and inquiry reasoning progressions that we mentioned earlier cannot be evaluated directly. Therefore, our item design included both forward and reverse engineering of assessment items to elicit a set of items that might provide information about students' knowledge development associated with each point of our content and inquiry reasoning progressions. Pretests and posttests were identical and evaluated students' content and reasoning over a range of complexities. The pre/posttest had a total of twenty-three items, with sixteen multiple choice/fill-in-the-blank items and seven open ended explanation items. Six items on the pretests and posttests were drawn from released standardized tests (two multiple choice items) from the Michigan Educational Assessment Program (http://www.michigan.gov/mde) and four items (two multiple choice, two open-ended explanations items) from the National Assessment of Educational Progress (http://nces.ed.gov/nationsreportcard) with the remaining 17 items written and pilot tested for the curriculum by the research team.

After pre/posttest development, we conducted several evaluations of our assessments. First, we explored the dimensionality of our items through a full information factor analysis (Thissen & Wainer, 2006) using ORDFAC that supports ordinal data (Schilling, 2002). Exploratory factor analyses indicated that the test items represented a unidimensional construct and were best fit by a single factor (Gotwals, 2006). All assessment items were initially mapped to a location on the biodiversity content and inquiry reasoning progressions; the mapping was later empirically evaluated through validity studies conducted prior to these experimental studies (Gotwals & Songer, 2006). We used Item Response Theory (IRT) (Hambleton, Swaminathan, & Rogers, 1991) in conjunction with WINSTEPS (Linacre, 2003) to create one parameter graded response models to investigate scale properties and to score students pre- and posttests in the respective measures. Our test curve information functions had IRT reliabilities of 0.78, 0.67, and 0.64 for all items, complex items and standardized items respectively. Our results demonstrated that all our items fit our model sufficiently; and inquiry reasoning-mapped categories of items (minimal, intermediate, and complex) in general corresponded to less, intermediate, and more complex item difficulty levels. Figure 2 provides an example of an item that maps to complex inquiry reasoning progression and the B3 location (including B1) of the content progression. For more information on item design and evaluation, see Gotwals and Songer (2006).

The first 8-week curricular unit included eight embedded assessments, with students completing approximately one embedded assessment per week. Embedded assessments were a regular part of the curricular activities. Embedded assessments mapped to a range of locations on the content progression but only one location, the intermediate location, of the inquiry reasoning progression. In other words, embedded assessment items consisted of scaffold-supported open-ended explanation items each of which corresponded to particular content topic and data set from the curricular unit. While the content hints focused on a different topic and different content progression mapping with each embedded assessment, the structural scaffolds (Reiser, 2004) guiding students to include the necessary components in their evidence-based explanations, remained consistent. This decision was made based on previous research project work demonstrating that in the early phases of the development of explanations, students require constant structural support (Lee and Smith, 1997).



A: Complex and B1/B3 Assessment Item

School Yard Animal Data:

Animal Name	Zone A	Zone B	Zone C	Total
Roly-poly	1	3	4	8
Ant	4	6	10	20
Robin	0	2	0	2
Squirrel	0	2	2	4
Pigeon	1	1	0	2
Animal Abundance	6	14	16	36
Animal Richness	3	5	3	5

Write a scientific explanation to the following question:

Scientific Question: Which zone most likely contains the most habitats?

B: Intermediate and E10 Assessment Item

Given the food chain: **Seeds → Mice → Snakes**

Scientific Question: What will happen to the food chain when there are a lot of seeds?
(Make sure that your explanation has a claim, 2 pieces of evidence, and reasoning)

Explanation:

Figure 2. Sample complex and intermediate assessment items.

Results from embedded assessments complement information obtained from pre/posttests in two ways. First, embedded assessments provide information on the nature and quality of explanations students can develop under guided conditions associated with a range of biodiversity topics and placements along the content progression. Second, embedded assessments provide more fine-grained information along a range of time points associated with progress during the curricular intervention.

Step 4: Research Studies to Evaluate Student Achievement

Perhaps the most essential step of our learning progression work is the empirical evaluation of our learning progression-developed products. This section outlines the evaluation of the 8-week curricular unit developed as a manifestation of the first content and inquiry reasoning progressions.

Sample

Our cohort was 1885 Detroit Public School sixth graders working with 22 DPS teachers in 18 research schools. Detroit students have nearly three times the average poverty rate of the state of Michigan and 92% are ethnic minorities. Sixth graders in Detroit are distributed across 35 schools; of these 18 schools were designated research schools while 17 were non-research schools. As illustrated in Figure 3, research students were not a privileged population of Detroit students as they scored lower than other students in Detroit on three of the four tests examined. Also of note is that the Detroit student average was significantly below the average for Michigan students reflecting an unfortunate and persistent trend nationwide of lower average performance by ethnic minority and poor students on standardized tests of this kind (Children's Defense Fund, 2006). Control students were students in research schools who throughout the same 8-week time period followed the district-approved, textbook-based curricular program focusing on biodiversity and ecology concepts. Both control and intervention program students took pre and posttests at the beginning and end of the 8-week time period.

Two types of research studies were conducted to gather empirical evidence of complex thinking about biodiversity associated with our sixth grade curricular unit. First, we implemented a cross-sectional investigation to provide information on the effectiveness of the biodiversity curricular program on student achievement through parallel hierarchical linear models (HLM). Second, we conducted growth curve analysis via a hierarchical piecewise linear growth model to descriptively examine students' complex biodiversity reasoning growth trajectories throughout the first curricular program.

Cross-Sectional Study

The cross-sectional study looked for empirical evidence associated with three measures of student achievement so as to compare the relative strength of each measure relative to one another. These measures were: (1) overall achievement including both biodiversity content and explanation development associated with

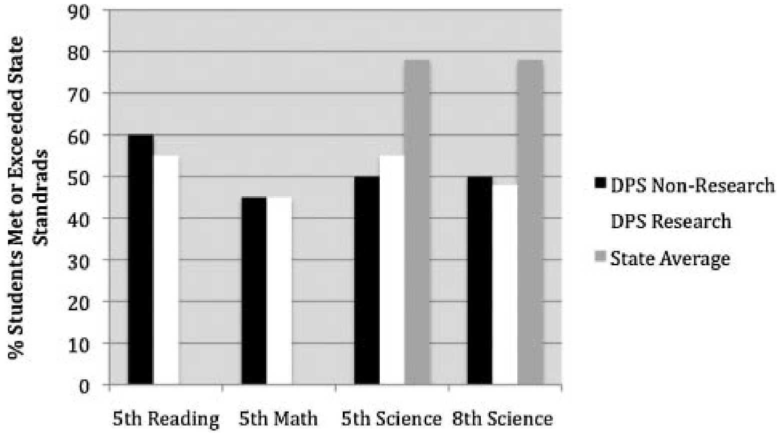


Figure 3. Average passing rates of Detroit Public School research and non-research students compared to state average on Michigan achievement tests.

the entire pre/posttest; (2) complex achievement indicating a subset of four of the seven open-ended test items that corresponded to the complex placement on the inquiry reasoning progression (Gotwals & Songer, 2006), and (3) standardized achievement indicating a subset of ten biodiversity items that were drawn from existing standardized tests and that mapped to several locations on the biodiversity content progression. All three dependent variables were continuous and approximately normally distributed measures of biodiversity achievement. Through student surveys, we also collected student measures primarily focused on addressing any pretreatment differences between students. The independent student variables focused on four general categories: (1) social background (racial status, sex, etc.), (2) prior academic history (pretest, grade repetition, etc.), (3) language (language spoken at home, length in US, etc.), and (4) home resources (computers, books, etc.). Teacher level covariates were derived from teacher interviews and teacher logs that were completed at least weekly. The treatment variable was a continuous measure of the percent of the intervention curricular program activities completed as recorded in the teacher logs. The values of this variable ranged from no intervention activities completed (control) to all activities completed with a mean of 0.58 (i.e., 58%) and SD of 0.32. Additionally, we utilized key teacher variables drawn from teacher interviews. Our class/teacher level measures represented typical characteristics and are summarized by eight categories: (1) years experience in teaching, (2) professional development, (3) experience with inquiry curricular programs, (4) human and physical resources, (5) classroom aggregates, (6) teacher confidence in teaching biodiversity, (7) reliable access to technology, and (8) academic background of the teacher. Table 2 presents the descriptive statistics for research teachers.

Missing Data. Rather than remove those students or teachers that have incomplete data, we employed the multiple imputation procedure to impute missing values (e.g., Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001). Although no data were missing on our teacher level variables, up to 30% of our student sample had at least one missing data point resulting from student mobility and/or absenteeism. Although unverifiable directly, our data suggest that student attrition was unrelated to achievement and program percent completed. Using multiple imputation, we generated five separate, multiply imputed, student level data sets. In an additional effort to increase the robustness of our inferences, we based the imputations on all available variables measured at both the student and teacher level (Peugh & Enders, 2004). Table 3 provides descriptive statistics on the raw and imputed data for student variables presented in subsequent analyses.

Analytical method. We recognize the multi-level nature of our data, for example, students are nested within classes/teachers. Accordingly, student characteristics are considered at level 1 and teacher characteristics are considered at level 2. Moreover, as we implemented the curricular program through the teachers and subsequently entire classes received an identical treatment dose, we considered the effect of the curricular program to be a level 2 treatment. We did not pursue a third level of the hierarchy nor fixed school effects since we have, on average, one to two teachers per school and thus cannot accurately partition the variance that is uniquely due to school

Table 2
Descriptive statistics for research school teachers (N = 22)

Teacher variables	Mean	SD
Proportion of teachers with undergraduate major of education	0.27	0.45
Proportion of teachers with undergraduate major of science	0.27	0.45
Proportion of teachers that regularly attend professional development	0.40	0.50
Proportion of classes with reliable access to technology	0.36	0.49
Number of inquiry curricula taught by teacher	3.50	1.79
Proportion of teachers certified in elementary science	0.68	0.47
Proportion of teachers certified in secondary science	0.27	0.45
Proportion of teachers with masters degree	0.68	0.47
Years of teaching experience	2.63	0.90
Proportion of teachers with high confidence in teaching biodiversity	0.59	0.50
Average class pretest score	37.51	2.19
Average class posttest score	40.94	2.69
Percent of class African-American	72.1	15.3
Percent of class Hispanic	11.9	21.8

Table 3
Descriptive statistics for students

	Raw data			Imputed data ($N = 1{,}885$)	
	N	Mean	SD	Mean	SD
Male	1,342	0.51	0.50	0.51	0.50
African-American	1,320	0.68	0.47	0.68	0.47
Hispanic	1,320	0.14	0.34	0.14	0.34
Lived in US less all of life	1,320	0.20	0.40	0.20	0.40
Non-English spoken at home	1,320	0.40	0.49	0.41	0.49
Have computer at home	1,320	0.71	0.45	0.71	0.45
BioKIDS pretest score	1,496	38.54	7.85	37.50	7.85
BioKIDS posttest score	1,425	42.28	7.45	41.03	7.17
Curriculum completion percent	1,885	0.58	0.32	0.58	0.32

characteristics. Although the number of groups at level 2 is small ($n = 22$), prior research has concluded that maximum likelihood in multilevel models with a small number of groups still provides unbiased estimates of fixed effects such as our treatment[1] (Browne & Draper, 2000; Van Der Leeden et al., 1997).

Model. In constructing our hierarchical model of achievement, we focus solely on a random intercept model (Appendix). Although hierarchical models allow within school (level 1) independent variables to vary randomly as well, we constrain these additional random effects to be zero as our primary interest rests on the average effect of the program. We centered all independent level 1 and level 2 variables around their respective grand means save the effect of the program and standardized the outcome (mean = 0, SD = 1). Our level 2 model exclusively models the average biodiversity achievement adjusted for students' academic and social backgrounds. With our fully unconditional model, we estimate approximately 13%, 12.6%, and 12.5% of the variance in overall biodiversity, complex, and standardized achievement, respectively, can be uniquely attributed to the teacher level (Table 4).

Results—Psychometrics. Our psychometric focus targets the amount of information our measures are extracting from students relative to each other. Psychometric analysis of the overall test indicated that this assessment provided a considerable amount of reliable information. The reliability of the test was 0.78 and it provided the maximum amount of information on students who are about three-fourth of a standard deviation below the mean student ability (Table 5). Although this measure is strongest when distinguishing among students with less than average ability, the test provides reliable and considerable information for students who are within ±2 and a 0.5 SD from the mean. Similar analyses demonstrated that complex and standardized items have a reliability of 0.67 and 0.64 and maximum information levels at 0.35 and −0.02, respectively. Embedded assessment psychometric properties are derived from a random sample of control students that were assessed under conditions comparable to that of the pre- and posttest. Embedded assessments were also well suited to the level of our students, with an IRT reliability of 0.81 and maximum amount of information on students who are 0.12 SD above the mean student ability. Therefore this assessment also provides reliable and considerable information and scores for students who are within ±2 and a 0.5 SD from the mean.

Figures 4 and 5 present psychometric data on the four measures. Figure 4 illustrates the characteristic curves of the measures. We see a much steeper slope (also see Table 5) for the complex assessment and the

Table 4
Fully unconditional model variance components and reliability

Variance component	All items	Complex items	Standardized items	Embedded tests[a]
Within classroom	0.87	0.88	0.87	0.30
Between classrooms	0.13	0.13	0.13	0.12
Intraclass correlation	13	12.6	12.5	0.28
Intercept reliability	0.9	0.9	0.9	0.76

[a] Variance components for embedded items are between time points and between students

Table 5
Psychometric properties of all measures

Measure	Reliability	Max information location	Slope
All items	0.78	−0.77	1.04
Complex	0.67	0.35	1.35
Standardized	0.64	−0.02	0.73
Embedded	0.81	0.12	1.50

flattest slope for the standardized assessment in the range surrounding average ability (i.e., 0). Evident from such slopes is the ability of the complex assessment scale to more reliably discriminate between students of differing abilities. Figure 5 provides information on the amount and range of information one test provides compared to others. These graphs indicate that the embedded assessment and the overall assessment provide the greatest amount and range of information, while the standardized test provides the least information. Although the overall assessment may benefit from having more items, the complex, standardized and embedded assessments had a similar number of items. The test information function for the complex assessment indicates that it provides the most information at approximately one-third of a standard deviation above the mean. The complex assessment's main contribution to the overall assessment is thus targeted toward above average abilities, and as a measure to discriminate between students who can demonstrate complex reasoning abilities or not (e.g., Fig. 4). In revisiting the intended role of the complex assessment subtest, we designed the items to focus on those tasks that require complex reasoning to extract information at above average levels. In contrasting this intended role to its realized role, we found strong evidence indicating that our complex assessment was indeed tapping into more complex phenomena. In contrast, though the standardized items are virtually centered at the average ability, these items extract less information in general and contribute comparatively little information to the overall assessment. In summary, though we temper conclusive claims and we intend to only use this information for subsequent developments, our empirical research provides supportive and informative evidence that our assessments are providing a greater range and amount of information that can more reliably discriminate between students of differing abilities than a contrasting standardized assessment measure which was also focused on biodiversity content.

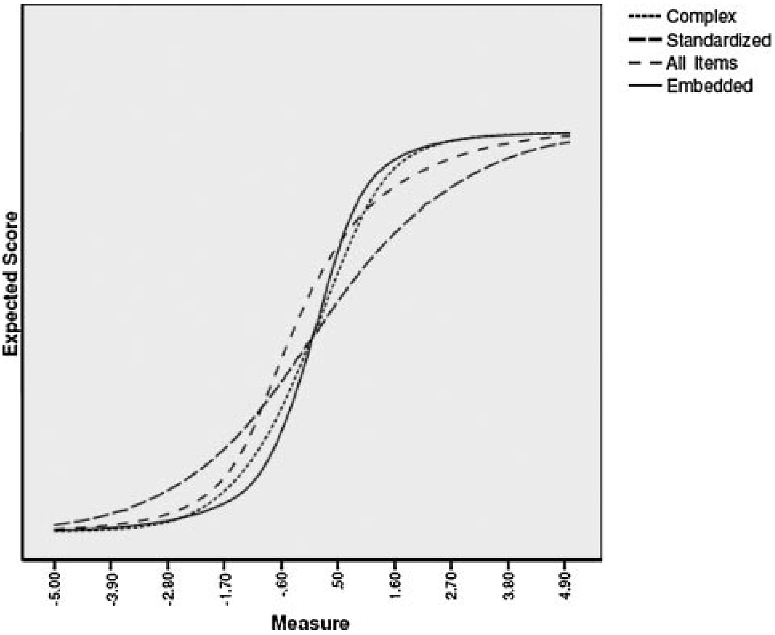


Figure 4. Item characteristic curves for complex, standardized, embedded and all items.

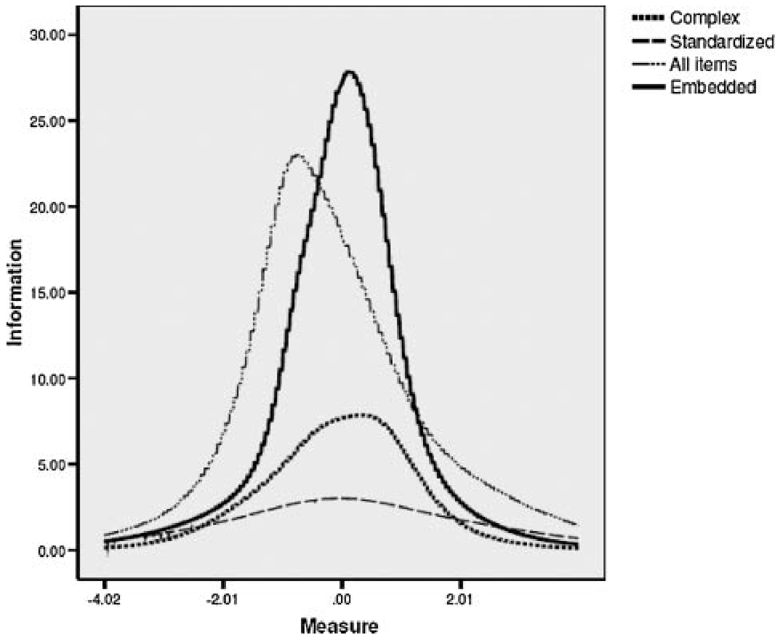


Figure 5. Item information curves for complex, standardized, embedded and all items.

Results—Achievement Models. Table 6 and Figure 6 summarize the achievement results from the cross-sectional models. The general model we selected considered student race, student English proficiency, student prior ability, class average prior ability, class percent of intervention program completed, teacher professional development and teacher confidence in teaching biodiversity. Our first model, concerning overall achievement, estimated the effect size associated with program completion to be a 0.34 SD ($p < 0.001$) gain when compared to students who did not participate in the program. Extending our results into complex and standardized sub-assessments (approximately one-third of the total items each), our models

Table 6
Biodiversity achievement

	General biodiversity achievement		Complex biodiversity achievement		Standardized biodiversity achievement	
Fixed effect	Coefficient	SE	Coefficient	SE	Coefficient	SE
Intercept	0.27**	0.08	0.10	0.8	0.253**	0.077
Average class pretest	0.03*	0.01	0.02	0.01	0.020	0.012
Intervention program	0.34***	0.11	0.62***	0.09	0.270*	0.111
Professional development	0.09	0.04	−0.01	0.06	0.099*	0.047
Confidence in teaching biodiversity	−0.08*	0.04	−0.06	0.04	−0.102**	0.032
Hispanic	0.26**	0.08	−.002	0.03	0.145	0.081
Language spoken at home	−0.06	0.03	−0.04	0.10	0.002	0.055
Pretest	0.18***	0.04	0.15***	0.04	0.342*	0.091
Multilevel program effect size	1.05		1.92		0.841	

Outcome is standardized (Mean = 0, SD = 1).
*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

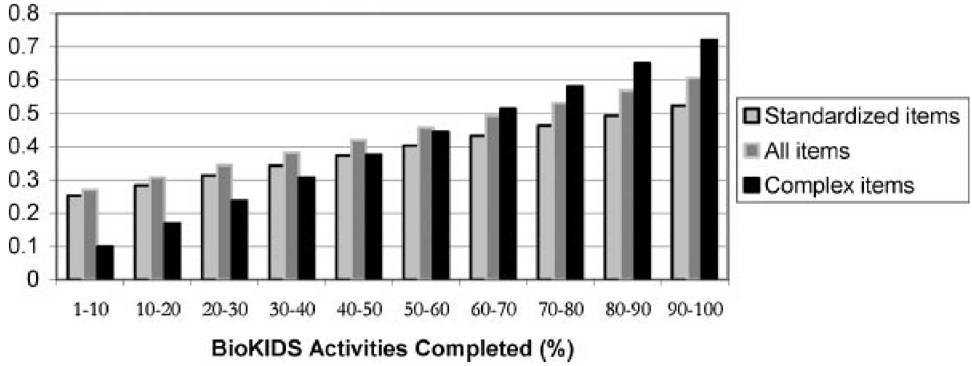


Figure 6. End of program achievement as a function of percent of intervention activities completed.

suggested that the program was associated with significant yet diverse effects. Specifically, we saw that achievement in intended target domains, for example, complex reasoning, was substantially higher (effect size of 0.62; $p < 0.001$) for students who completed our program. As anticipated, intervention students' achievement on standardized measures was also significantly better than traditional curriculum (control) students and had an effect size of 0.27 ($p < 0.05$). We present our effects in standardized effect sizes and interpret them as holding all other factors constant, a student in a classroom completing the intervention program will gain, on average, 0.34 SD more than a student who is in a classroom with a traditional curriculum. We note that our estimates are based on interval scales, multiply imputed data, and account for clustering. When comparing these results with, for example, a naïve model that uses listwise deletion for students without complete data rather than multiple imputation, our model of overall achievement indicates a program effect size of 1.10. This naïve estimate is more than three times the effect size of our adjusted estimate of 0.34 and is further amplified if we fail to account for clustering through an HLM.

Figure 6 contrasts the achievement results for the standardized, overall and complex assessments for students exposed to the program. In contrasting the program effects, we observed that overall achievement and achievement on standardized measures tends to be substantial (approximately one-fourth of a standard deviation—see intercepts) for both intervention students and non-intervention students. In other words, there is a difference, but not a substantial difference, between intervention and non-intervention students on knowledge as measured by standardized test items. Conversely, student achievement on complex reasoning tasks is highly dependent on how much of the intervention curriculum a student has completed. Students who have no involvement with the program gain, on average, only one-tenth of a standard deviation on complex tasks throughout the 8 weeks. Students who are exposed to the intervention start with small gains but demonstrate sizeable gains by the end of the program. These results suggest that our measure of complex reasoning is a slow developing ability that needs cultivation over an extended and focused period.

Growth Curve Study

In order to extend our cross-sectional analyses and link individual student growth with their respective characteristics, we developed a model that explores students' explanations about biodiversity growth trajectories throughout the program. We wanted to inform our curricular development by examining the nature and rate of student growth in complex reasoning on a week-by-week basis throughout our curricular unit. Though we accept that pre–posttest achievement is well approximated by a linear trajectory, our weekly progress assessments allow us to use a piecewise growth model to estimate nonlinear growth.

Preceding program implementation, we selected a purposeful sample ($n = 6$) of the original 22 teachers and their respective students based on high intended intervention completion percent and their diverse range of student characteristics. Growth curve data included eight embedded assessments each from 567 students associated with 6 teachers. As mentioned earlier, each embedded assessment focused on an intermediate inquiry reasoning placement paired with a different content progression placement. We made use of the sample's weekly progress and their nested nature (multiple time points nested in students) by modeling the trajectories as a piecewise hierarchical growth model that is linear between contiguous time points/activities. Level 1 focuses on how students grow over time, whereas the second level explores how the individual characteristics interact with specific portions of our program (Appendix). Although two additional levels of nesting exist (students nested in classrooms which in turn are nested in schools), our interest (and data) here is in student's interactions with the program.

Analogous to the previous cross-sectional models, our time invariant (level 2) student measures were derived from the student survey and focused on the same categories. Moreover, similar to the first set of analyses, we utilize multiple imputation to handle missing data (approximately 20%). The growth curve sub sample (Table 7) differed from the cross-sectional sample (Table 3) in several aspects. Briefly, the growth curve sample tended to have more male students, less African-American students, more Hispanic students, more students who families spoke another language at home and higher pretest scores. Therefore, we do not attempt to generalize this sub sample to our larger sample but rather utilize the growth model to provide empirical descriptions of our programs strengths and weaknesses.

In an effort to establish and adjust for the varying difficulties of the embedded assessments, we investigated the psychometric properties using a random sample of students from control schools. We administered all eight embedded assessments to this random sample. We created a one parameter graded response model using the anchored difficulty weights from the control sample to assign students biodiversity reasoning ability scores for each time point. Our model suggested our embedded assessments targeted students near average ability and had an IRT reliability of 0.81.

Model. We built the growth curve model in progressive stages by examining the student categories sequentially. Moreover, we retained (in all level 2 equations) variables that were statistically related to the outcome in any of the level 2 equations and built each equation with the same set of predictors to help mitigate possible dependencies between fixed effects for the intercept and slope (Raudenbush & Bryk, 2002). We first examined prior academic history variables, followed by a consideration of social background covariates in conjunction to academic history. Subsequent student categories were constructed only after settling on the final set of measures for previous categories.

Level 1 of our HLM piecewise growth model is the model for individual growth and represents the change in complex reasoning during the specified time period. As we hypothesized that growth may not be linear and may occur in a variety of spurts, we used a piecewise linear model which allows different growth slopes between any two successive time points. Level 2 examines how students with various characteristics interact with specific components of the program by specifying each growth period function of time-invariant student characteristics. For each time point, we modeled its coefficient as a function of race, English proficiency, and prior ability.

Results—Growth Curve Model. Figures 7 and 8 present student growth curves for the eight embedded assessment tasks by race and pretest quartile, respectively. As projected, students' of all sub populations show substantial growth, and the pattern of growth shows evidence of growth spurts and plateaus. These data illustrate the largest growth spurt between the start of the program and the end of week 3 with diminished growth rates throughout the remainder of the program. While evidence from the posttest achievement data

Table 7
Descriptive statistics for growth curve students (N = 567)

	M	SD
Proportion male	0.57	0.50
African-American	0.47	0.50
Hispanic	0.32	0.46
Lived in US *less* all of life	0.25	0.43
Non-English spoken at home	0.54	0.50
Have computer at home	0.69	0.47
BioKIDS pretest score	38.30	6.58

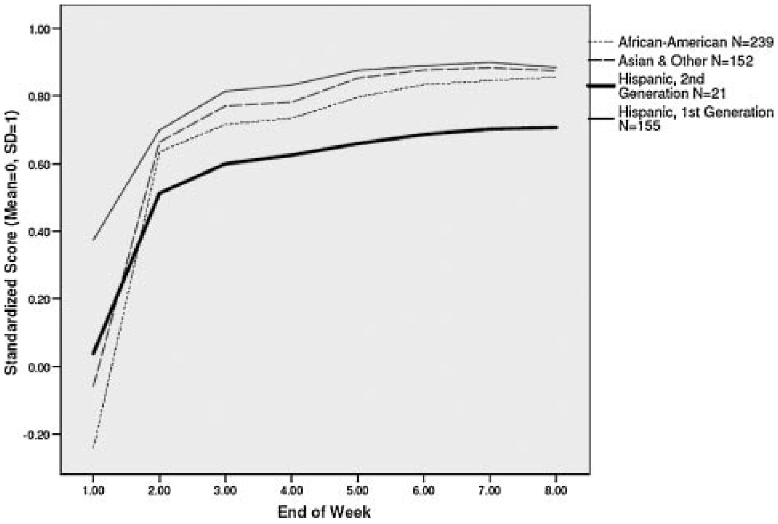


Figure 7. Complex biodiversity reasoning growth throughout program for African-American, Asian and Other, Hispanic, Second Generation and Hispanic, First Generation Students.

(Fig. 6) provides evidence that intervention students continue to grow in their complex abilities through our program, the ability of our embedded assessments to detect growth becomes nearly saturated by the end of the third week. This result corresponds with our test information function for embedded assessments as we see the assessments ability to extract information on students ability becomes increasingly limited as students grow. Explicitly, these results provide important information to guide improvement of our first version of assessment design: weeks 4–8 assessments need to involve more advanced levels of complexity in order to adequately measure student's ability and growth in the later time points. Implicitly, these results also provide suggestions for possible curriculum and learning progression redesign: they suggest that we may more effectively cultivate complex reasoning by removing curricular scaffolds or increasing complexity at earlier time points.

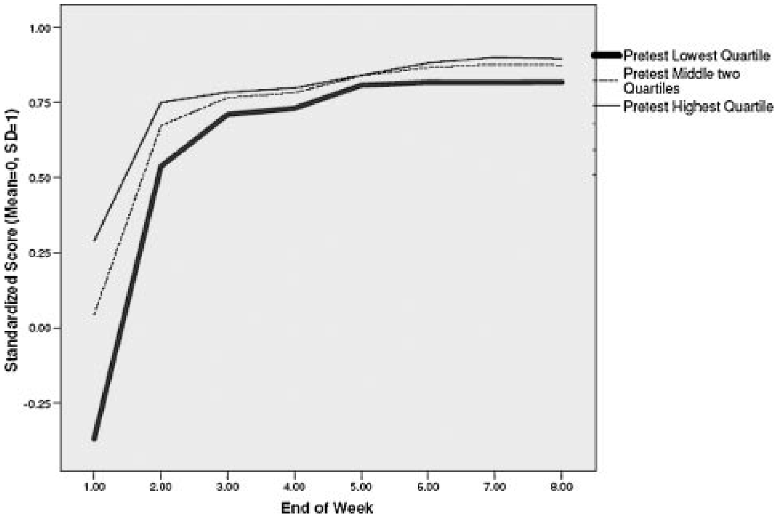


Figure 8. Complex biodiversity reasoning growth by pretest ability.

Results—Sensitivity Analyses. Our estimates for the complex, standardized and overall batteries are valid and statistically significant only if a number of assumptions are met. Two such conditions are the validity of our HLM standard errors given our level 2 sample size and the assumption that there is not an unmeasured variable confounded with the treatment exposure. In this section we evaluate the robustness of our model-based estimates to both of these assumptions. We consider our effects sensitive if the new confidence interval includes zero (or $p > 0.05$).

The statistically significant association between student achievement and our program is based on our model's standard errors. However, in multilevel settings where the number of groups is limited, though fixed effects are unbiased, standard errors tend to have a downward bias (see Note 1). Our first set of sensitivity analyses examines how our inferences might change had the model-based standard errors been inflated by a large amount, say 50%. The statistical significance of our estimates of the overall effect and complex reasoning effect change slightly but remain robust ($p < 0.05$) to inflated standard errors of this magnitude. However, our estimates of the standardized achievement effect are sensitive and the inflated confidence interval includes zero ($p = 0.12$). Table 7 displays our adjusted estimates.

Our estimates are unbiased only if our assumption that there is no unmeasured variable that is confounded with treatment assignment given the observed variables is defensible. Although we can attempt to measure and control for confounding variables such as prior ability, it is not possible to exhaustively measure or control for every potentially confounding variable. As a result we are forced to assume that any unmeasured covariates are independent of treatment assignment given our measured covariates.

We assessed the robustness of our associational inferences to the inclusion of an unmeasured variable via a sensitivity analysis (Hong & Raudenbush, 2006; Rosenbaum, 1986). This analysis constructs a sensitivity index from the set of observed measures to determine if our model predictions are significantly influenced by potential hidden biases resulting from unobserved covariates. In particular, we examine the potential bias resulting from unequal assignment to treatment levels. In our sensitivity analysis, for all three effects, we examined whether our estimates would be significantly altered by additional adjustments for an unmeasured confounder. We examined the impact of omitting one of the measured potential confounders on the estimates of the treatment coefficient and our inferences. We use these adjusted estimates as an index to gauge the robustness of the program effect and regarded each effect as sensitive to an unmeasured confounding variable if the new estimate was not significantly different from zero (i.e., $p > 0.05$).

Our sensitivity analyses indicated that our estimates of the program effect on overall, complex and standardized assessment are insensitive to an omitted variable similar to those measured (Table 8). In our sample, the percent of students who are Hispanic demonstrated the strongest relationship to treatment level in all three assessments whereas the average pretest ability of the classroom illustrated the strongest relationship with all three outcomes. Combining these two relationships to form a hypothetical unmeasured confounding variable, we estimated the adjusted effect.

Though our original model-based estimates are insensitive to a hypothetical confounder, we note that our sensitivity analyses consider a limited set of covariates. In particular, common support between treatment

Table 8
Effects after controlling for a hypothetical confounding variable

Measure	Original effect	Effect adjusted for confounder[a]	Effect adjusted for inflated standard errors[b]
Overall	0.34***	0.30**	0.34*
Standardized	0.27*	0.24*	0.27 NS
Complex	0.62***	0.58***	0.62***

NS, not significant ($p > 0.05$).

[a] Effect adjusted for hypothetical confounder constructed from Percent Hispanic and average class pretest ability.

[b] Statistical significance adjusted for inflated (by 50%) standard errors.

*$p < 0.05$.

**$p < 0.01$.

***$p < 0.001$.

levels is not complete and thus counterfactual estimates are partially extrapolated. Contrastingly, there is strong evidence that the inclusion of additional covariates accounts for decreasing amounts of variance once the most predictive are considered (Bloom, 2005). In the current context, variables such as prior ability, race, teacher professional and educational backgrounds as well as others measured are strongly associated with student achievement. Including these variables suggests that they would be correlated with any unmeasured confounders and thus absorb or reduce the impact of an unmeasured confounding variable (Frank, Duong, Maurolis, & Kelcey, 2008).

Step 5: The Revision and Expansion of Learning Progressions into 3-Year Sequences

The final step in our design process was to build from the empirical evaluation of the first work products and the national and state science standards to develop the first 3-year learning content and inquiry reasoning progressions. Table 9A presents our 3-year content progression for biodiversity. This content progression may appear to represent more rather than fewer content topics. In actuality, the grain size of our content progression focal points is much smaller than topics in a traditional curricula (e.g., Lisowski & Jones, 2007).

Table 9A
Three-year content progression for biodiversity

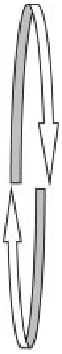
6th grade			B11: Human activity and other factors affect biodiversity of ecosystems (introduced species, changing habitat qualities, food web disruptions). ... B7: Biodiversity differs in different areas. It is a useful way of characterizing habitats
		E10: Because many animals rely on each other, a change in the number of one species can affect many different members of the web. ... E7: You can connect the plants and animals of a habitat into a food web	
5th grade			B5: An area has a high biodiversity if it has both high richness and abundance ... B1: A habitat is a place that provides food, water, shelter and space for living things
	C8: Patterns of shared characteristics reveal the evolutionary history of groups ... C5: Organisms are grouped based on the structures they have in common		
4th grade		E6: Only a small fraction of energy at each level of a food chain is transferred to the next level ... E1: Every organism needs energy to live and gets that energy from food	
	C4: Organisms have different features that allow them to survive. ... C1: There are observable features of living things		
	Classification	**Ecology**	**Biodiversity**

Comparison of our focal points to topics from the district approved textbook for the same material (Lisowski & Jones, 2007) reveals that our 3-year program addresses three of 20 comparable topics (classification, ecology and biodiversity), and provides much more focus associated with each topic. This focus on fewer topics supports the general assumption in learning progressions work that fewer topics will be addressed sequentially, revisited, and addressed in greater depth over multiple years as compared to a presentation that might occur with a less organized 3-year sequence or a more declarative-knowledge based approach common in many textbook-based science programs.

Achievement results, particularly our results on complex items, and embedded assessments suggested that students needed systematic and more complex content, both within one curricular unit and across the 3-year sequence. These results guided the development of a content progression that sequenced content around three interrelated strands C: Classification, E: Ecology and B: Biodiversity. Table 9A illustrates our expanded content progression that summarizes our first 3-year template of this kind. In this content progression, we begin with four classification ideas in 4th grade, then build on those with an additional four classification ideas in 5th grade. Ecology is introduced in 4th grade, and these ecology ideas serve as a foundation for additional ideas in ecology in 6th grade. Biodiversity is introduced in 5th grade and built upon in 6th grade.

Embedded assessment results also provided suggestions on how to organize our inquiry reasoning progression and associated curricular products. Table 9B presents our 3-year inquiry reasoning progression associated with building evidence-based explanations. Building from our results we made two changes: First, our revised inquiry reasoning progression allows scaffold support on the components of an evidence-based explanation, such as what constitutes evidence (level 2s), prior to support on the construction of a complete explanation (level 4s). We speculate that this approach will facilitate both more guidance on the components of an explanation associated with different data and contexts, as well as more information on what kinds of knowledge students are lacking in their development of evidence-based explanations over a range of biodiversity sub-topics and situations. Second, we speculate that unlike our content progression that we expect to be transformed into curricular materials in a linear fashion (e.g., content sequence fourth grade focal points are manifested into fourth grade activities and fifth grade focal points are manifested into fifth grade activities), we expect our revised inquiry reasoning progression might be best manifested into curricular activities in a cyclical manner. For example, fourth grade curricular activities might manifest C1–C4 content at scaffold levels (1s, 2s, 3s, 4s) followed by E1–E6 content also at scaffold levels (1s, 2s, 3s, and 4s). In fifth and sixth grade when Classification and Ecology topics are revisited and built upon, students would pick up with C5 or E6 content through scaffold activities (1s–4s) that quickly progress to unscaffold levels with this content material (1, 2, 3, and 4). Our work is currently in progress to develop curricular products that manifest these ideas and which can be used to empirically test these hypotheses.

Table 9B
Three-year inquiry reasoning progression for building evidence-based explanations



Level 4: Student constructs a complete scientific explanation (*without* scaffolding)

Level 4s: Student constructs a complete scientific explanation (*with* scaffolding)

Level 3: Student makes a claim, backs it up with evidence, and provides reasoning to tie the two together (*without* scaffolding)

Level 3s: Student makes a claim, backs it with evidence, and provides reasoning to tie the two together (*with* scaffolding)

Level 2: Student makes a claim and backs it with sufficient and appropriate evidence (*without* scaffolding)

Level 2s: Student makes a claim and backs it with sufficient and appropriate evidence (*with* scaffolding)

Level 1: Student makes a claim (*without* scaffolding)

Level 1s:Student makes a claim (*with* scaffolding)

Conclusions

Recognizing a need to guide students to develop essential complex thinking about important science topics, the approach of learning progressions was developed to guide the organized creation of curricular programs to foster sophisticated thinking about essential topics over multiple units and years. Our focus in this article was to implement a process of learning progression template development, empirical evaluation and product refinement. Following a five step process, we developed first versions of learning progressions, used these templates to develop curricular and assessment products, then utilized these products to gather empirical information to guide the development of 3-year progressions to be used in additional rounds of product development and evaluation. We adopted an empirically driven, five-step process to investigate what sequence of concepts and inquiry reasoning skills foster complex thinking about biodiversity. In addition, these studies were designed to gather empirical evidence on student achievement of complex thinking about biodiversity as measured by different assessment instruments focused on biodiversity content including standardized tests, embedded assessments, and instruments designed to measure complex thinking about biodiversity.

Our work in assessment development provided evidence that learning progression-guided assessments provide a greater range and amount of information that can more reliably discriminate between students of differing abilities than a contrasting standardized assessment measure that was also focused on biodiversity content. Applying new approaches in educational statistics, our studies demonstrated that our assessment instruments provided reliable and considerable information for our target population, and our assessments provided more information on complex reasoning than a comparable standardized test instrument. While intervention and control students both made significant gains, and similar gains, on standardized items, intervention students demonstrated substantial achievement gains as compared to control students on complex reasoning tasks.

While we value our achievement results, we believe our work in assessment to illustrate the insensitivity of standardized tests to evaluate complex thinking about science is perhaps the most important aspect of this work. Figure 6 illustrates that if we had utilized only a standardized test instrument to evaluate program effects, we would not have nearly as much information on the character and amount of improvement students demonstrated on complex reasoning tasks. Results of this kind point to the need for learning progressions to not only guide the development of curricular products, but also guide the development of assessment instruments to ensure the most and most accurate information can be gathered associated with the evaluation of learning progression-related products.

In addition, our research efforts resulted in other developments. First, we chose to conduct both achievement and growth curve analyses in order to intentionally obtain a range of empirical information on the evaluation of our first version of products. Growth Model research results demonstrated evidence of both strengths and weaknesses of the first version of our curricular activities via the mapping of students' growth trajectories throughout the unit. Comparing results from achievement and growth curve results provide interesting hypotheses, such as while more time with the program continues to advance students general understanding and understanding on complex items (e.g., Fig. 6), embedded assessments suggest an overall weakness in this unit in regards to maximally fostering high growth rates throughout the entire 8-week period. We continue to believe that gathering selected contrasting empirical information allows us to make more informed decisions than either empirical approach alone.

Second, we intentionally applied new developments in statistical analyses towards the evaluation of learning progression products in order to not only evaluate our products, but to advance our understanding of best empirical evaluation approaches. While these analytical approaches are only a few of the possible means of evaluating the development of learning progressions products, we encourage similar efforts to not only evaluate products but to advance understanding of best means for information gathering.

Collectively, we believe that these kinds of data and research approaches can provide us with valuable information on how our learning progression-guided program interacts with students and how those interactions manifest into growth spurts or growth plateaus. While we have as yet only conducted the first round of our empirical testing of learning progression products, we hope research and development in this important area of learning progressions can advance our collective understanding of both achievement in science, the strengths and weaknesses of existing standardized tests, and the character of "middle knowledge"; in other words the varieties of not-quite-successful attempts at complex scientific ideas that students manifest on the path towards sophisticated understanding. We encourage additional work that continues to challenge existing thinking not only of learning progression-associated learning outcomes and their implications, but the application of new research and analytical approaches that can increase the range and quality of empirical evidence available for sound decision-making.

This material is based in part upon research supported by the National Science Foundation under grants REC-0089283 and REC-0129331. Any opinions, findings, and conclusion or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors gratefully acknowledge Philip Myers, the entire BioKIDS research team, and the Detroit Public School teachers and students for their support.

Notes

[1] In addition the standard errors tend to be downwardly biased. With the number of groups in the 20s, the type one error rate at the alpha level of 0.05 tend to be inflated to only about 0.09 (Browne & Draper, 2000). Moreover, estimation via restricted maximum likelihood compared with full maximum likelihood provides even more reliable variance estimates (Van Der Leeden, Busing, & Meijer, 1997). This is subsequently addressed in sensitivity analyses.

References

Bloom, H. (2005). Randomizing groups to evaluate placed-based programs. In: H.S. Bloom (Ed.), Learning more from social experiments: Evolving analytic approaches (pp. 115–172). New York, NY: Russell Sage Foundation.

Browne, W.J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. Computational Statistics, 15, 391–420.

Bruner, J. (1996). The culture of education. Cambridge, MA: Harvard.

Catley, K., Lehrer, R., & Reiser, B., (2004) Tracing a prospective learning progression for developing understanding of evolution. Paper commissioned by the National Academies Committee on Test Design for K-12 Science Achievement.

Children's Defense Fund. (2006). *State of America's Children 2005*. Retrieved January 5, 2007, from http://campaign.childrensdefense.org/publications/greenbook/default.aspx.

Frank, K., Duong, M., Maurolis, S., & Kelcey, B. (2008). Quantifying statistical control: Crossing the threshold of randomization. Paper presented at the Society for Research on Educational Effectiveness.

Gotwals, A.W. (2006). *The nature of students' science knowledge bases: Using assessment to paint a picture*. Unpublished Doctoral Dissertation. Ann Arbor, MI: The University of Michigan.

Gotwals, A.W., & Songer, N.B. (2006). Measuring students' scientific content and inquiry reasoning. Proceedings of the 7th International Conference of the Learning Sciences. Bloomington, IN.

Hambleton, R., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of item response theory. London: Sage.

Hong, G., & Raudenbush, S. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. Journal of the American Statistical Association, 101(475), 901–910.

Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York, NY: Basic Books.

Karplus, R., & Their, H.D. (1967). A new look at elementary school science: Science curriculum improvement study. Chicago, IL: Rand McNally.

Lee, V., & Smith, J. (1997). High school size: Which works best and for whom. Educational Evaluation and Policy Analysis, 19(3), 205–227.

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

Linacre, J.M. (2003). WINSTEPS Rasch measurement computer program. Chicago, IL: Winsteps.com.

Lisowski, M., & Jones, L.C. (2007). Environmental science. Boston, MA: Pearson.

Metz, K. (2000). Young children's inquiry in biology: Building the knowledge bases to empower independent inquiry. In: J. Minstrell & E. van Zee (Eds.), Inquiring into inquiry learning and teaching in science (pp. 371–404). Washington, DC: AAAS.

Mislevy, R.J., Almond, R.G., & Lukas, J.F., (2004) *A Brief Introduction to Evidence-Centered Design (Technical)*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST) UCLA.

National Academy of Science. (2007). Rising above the gathering storm: Energizing and employing America for a brighter economic future. Washington, DC: National Academies Press.

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

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

OECD. (2007). PISA 2006: Science Competencies for Tomorrow's World, Vol. 1: Analysis. Paris, France: Organisation for Economic Co-operation and Development.

Peterson, A.T. (2003). Predicting the geography of species' invasions via ecological niche modeling. Quarterly Review of Biology, 78, 419–433.

Peugh, J.L., & Enders, C.K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of Educational Research, 74(4), 525–556.

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

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.

Raudenbush, S.W., & Bryk, A.S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd edition). Thousand Oaks, CA: Sage.

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

Rosenbaum, P. (1986). Dropping out of high school in the United States: An observational study. Journal of Educational Statistics, 11(3), 207–224.

Schilling, S. (2002). ORDFAC: Ordinal factor analysis software. Ann Arbor, MI: University of Michigan.

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., Huber, A.E., Adams, K., Chang, H.Y., Lee, H.S., & Jones, T. (2005). BioKIDS: Kids' inquiry of diverse species, an eight-week inquiry curriculum using simple, powerful technologies. Ann Arbor, MI: The University of Michigan.

Thissen D., & Wainer H. (Eds.) (2001). Test scoring. Mahwah, NJ: Lawrence Erlbaum.

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

Van der Leeden, R., Busing, F., & Meijer, E. (1997). Bootstrap methods for two-level models, technical report no. PRM 97-04. Leiden, The Netherlands: Leiden University.

Appendix: COMPLEX REASONING ACHIEVEMENT MODELS AND GROWTH CURVE MODELS

Achievement Models

$$\text{Level 1}: \quad Y_{ij} = \beta_{0j} + \sum_{p=1}^{q_x} \beta_{ij} X_{pij} + \epsilon_{ij} \text{ and } \epsilon_{ij} \tilde{N}(0, \sigma^2)$$

$$\text{Level 2}: \quad \beta_0 = \gamma_{00} + \gamma_{01} Z_j + \left(\sum_{q=1}^{Q} \gamma_{0q} W_j \right) + u_{0j} \text{ and } u_{0j} \tilde{N}(0, \tau)$$

Growth Curve Models

$$\text{Level 1}: \quad Y_{ij} = \pi_{0j} + \sum_{p=1}^{q_x} \pi_{ij} A_{pij} + \epsilon_{ij} \text{ and } \epsilon_{ij} \tilde{N}(0, \sigma^2)$$

$$\text{Level 2}: \quad \pi_0 = \beta_{00} + (\sum_{q=1}^{Q} \beta_{0q} X_j) + u_{0j} \text{ and } u_{0j} \tilde{N}(0, \tau)$$

$$\pi_p = \beta_{p0} + \left(\sum_{q=1}^{Q} \beta_{pq} X_j \right)$$

where *A* represents time points, *X* represents student variables, and *W* represents teacher variables. The Growth Curve Models use the first time point (end of first week) as the reference point, and it summarizes the growth trajectories between adjacent time points through a linear approximation represented by a separate growth parameter ($\boldsymbol{\pi}_i$) for each period.