JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 50, NO. 6, PP. 639–659 (2013)

Research Article

Analyzing Change in Students' Gene-to-Evolution Models in College-Level Introductory Biology

Joseph T. Dauer,[1] Jennifer L. Momsen,[2] Elena Bray Speth,[3] Sasha C. Makohon-Moore,[4] and Tammy M. Long[1]

[1]*Department of Plant Biology, Michigan State University, 612 Wilson Rd. S-340, East Lansing, Michigan 48824*
[2]*Department of Biological Sciences, North Dakota State University, Fargo, North Dakota*
[3]*Department of Biology, Saint Louis University, Saint Louis, Missouri*
[4]*Program in Cell and Molecular Biology, Duke University, Durham, North Carolina*

Received 22 February 2013; Accepted 31 May 2013

Abstract: Research in contemporary biology has become increasingly complex and organized around understanding biological processes in the context of systems. To better reflect the ways of thinking required for learning about systems, we developed and implemented a pedagogical approach using box-and-arrow models (similar to concept maps) as a foundational tool for instruction and assessment in an Introductory Biology course on genetics, evolution, and ecology. Over the course of one semester, students iteratively constructed, evaluated, and revised "Gene-to-Evolution" (GtE) models intended to promote understanding about the connections linking molecular-level processes with population-level outcomes. In practice, a successful GtE model contextualizes information provided in a case study and explains how genetic-level variation originates at the molecular level, is differentially expressed among individuals, and is acted upon by the environment to produce evolutionary change within a population. Our analyses revealed that students' ability to construct biologically accurate models increased throughout the semester. Model complexity peaked near mid-term then subsequently declined. This suggests that, with time, students were building more parsimonious models, shedding irrelevant information, and improving in their ability to apply accurate and appropriate biological language to explain relationships among concepts. Importantly, we observed the greatest relative gains in model correctness among students who entered the course with lower mean GPA. In an analysis comparing performance among achievement tritiles, lower-performing students effectively closed the achievement gap with the highest performing students by the end of the semester. Our findings support the effectiveness of model-based pedagogies for science teaching and learning, and offer a perspective on pedagogical application of modeling strategies to foster systems thinking and knowledge structuring in college-level biology. © 2013 Wiley Periodicals, Inc. J Res Sci Teach 50: 639–659, 2013

Keywords: conceptual models; model-based pedagogy; assessment; undergraduate education; evolution

Momentum is growing among biology educators for rethinking our approaches to teaching and learning in college-level biology. In particular, there is interest in employing strategies that help students transition toward a way of thinking that is more reflective of contemporary

Correspondence to: Tammy M. Long; E-mail: longta@msu.edu
DOI 10.1002/tea.21094
Published online 25 June 2013 in Wiley Online Library (wileyonlinelibrary.com).

biology—one that deemphasizes fact accrual in favor of connecting processes that comprise biological systems. Systems are a crosscutting concept in the new US national science framework for primary and secondary education (National Research Council, 2012) and one of the core concepts suggested for postsecondary biology education (AAAS, 2009). The NRC framework and AAAS report also promote model construction and interpretation as ways to engage students in practices that mirror biological research. In our work, we explore the potential of iterative conceptual modeling of biological systems as an instructional tool to help students organize knowledge while revealing change in students' thinking during a semester of instruction in college-level Introductory Biology.

For students majoring in life sciences, Introductory Biology is a gateway to upper-division courses required for their majors, but can also prove challenging for retention in the discipline (Haak, HilleRisLambers, Pitre, & Freeman, 2011; Mervis, 2011). Attempts to cover topics ranging from DNA structure and function to conservation biology can leave students feeling overwhelmed by facts and unable to formulate big picture understandings that explain how subjects are connected. In a reformed, learner-centered version of Introductory Biology, we incorporated conceptual modeling as a core activity and assessment intended to facilitate students' abilities to make connections among biological concepts, for example connecting the principles of genetics and evolution. This approach departs from traditional instruction that teaches DNA structure–function and mutation as "cell and molecular biology" and separately teaches inheritance and evolution in the "organismal" and "population" units, respectively. Rather, we aimed to fully integrate students' learning across scales, where molecular-level processes serve as explanatory tools accounting for phenomena we can observe and quantify at the population level. In particular, we sought to progressively develop students' understanding about principles and processes of genetics that explain evolutionary outcomes, or "gene-to-evolution" (GtE) thinking.

To develop students' GtE thinking, we incorporated frequent, iterative modeling activities that engaged students in thinking about how molecular-level phenomena are manifest as organismal- and population-level outcomes. Throughout a semester of instruction, students constructed progressively more comprehensive models that represented (1) how random mutations arise in a genome, (2) how variation at the molecular level has consequences for traits expressed by organisms, and (3) how resultant variation among organisms can interact with environmental variables, resulting in population-level change.

Developing a GtE model poses several challenges, as it requires articulating relationships within and across spatial and temporal scales. Students need to reason about events that occur at the molecular level (gene mutation and genetic recombination), cellular level (transcription and protein synthesis), organismal level (phenotypic expression, differential fitness), and population level (allele frequency change over time). Conveying cross-scale dynamic interactions is cognitively very demanding (Feltovich, Coulson, & Spiro, 2001; Hmelo-Silver, Marathe, & Liu, 2007). Genetics and evolution are independently challenging sub-disciplines of biology, each well described in the literature as being conceptually difficult for learners. Learning genetics, for example, requires acquisition of a complex vocabulary, mathematical reasoning, and multi-level thinking (Bahar, Johnstone, & Hansell, 1999; Marbach-Ad & Stavy, 2000). Morphological characteristics of organisms (phenotypes) are generally observable at the "macroscopic" level, but they are due to genes and alleles, which are "microscopic" structures. Alleles and genotypes, in addition, are often represented by letters, the "symbolic" level (Johnstone, 1991). While experts easily incorporate all three levels in their genetics reasoning, students tend to focus on one level at a time and struggle to merge the three (Bahar et al., 1999). Students simultaneously learn concepts at different levels of biological organization but fail to grasp the interconnections among them (Marbach-Ad & Stavy, 2000).

Students encounter numerous barriers when learning about evolution as well, and these are among the best-characterized pedagogical challenges in the science education literature, including a variety of cultural, and cognitive constraints (Smith, 2010a, 2010b). Insights from the cognitive sciences suggest that understanding evolution is difficult, in part, because it requires a substantial amount of conceptual change, as learners typically need to accommodate new information that conflicts with existing robust naïve conceptions (Sinatra, Brem, & Evans, 2008). Consequently, students' evolutionary explanations are typically incomplete and consist of a mix of correct key concepts, naïve ideas, and cognitive biases (Nehm & Reilly, 2007; Nehm & Schonfeld, 2008). Solving evolutionary problems poses daunting challenges for novice biology learners, as it requires considering context, the role of emergent properties, and proximate and ultimate causal explanations (Nehm & Ridgway, 2011). Among seven key concepts that comprise a complete and accurate explanation of evolution by natural selection (Nehm & Schonfeld, 2008), two concepts belong to the domain of molecular and cellular genetics (the origin of variation and heredity) and a third is in the domain of population genetics (allele frequency change over time). In an evolution problem-solving study, most experts consistently incorporated the genetic causes of variation and heredity as key concepts in their evolutionary explanations; only a minimal percentage of students, however, included these concepts across their answers to the same problems (Nehm & Ridgway, 2011). Consistent with this result, in a previous iteration of the course described in this study, we observed that Introductory Biology students' written explanations of natural selection were largely centered on phenotypic variation and differential survival, but did not include sufficient details or relationships to adequately account for the genetic basis of evolution (Bray-Speth, Long, Pennock, & Ebert-May, 2009). Specifically, most students did not recognize how variation within populations needed to originate at the genetic level and to be heritable, to serve as the basis for evolution. Despite the evident centrality of genetics to a complete understanding of evolution, the two topics are typically compartmentalized in textbooks (Nehm et al., 2009) and often taught in isolation of one another, even in Introductory Biology sequences.

Schema and the Representation of Knowledge

One interpretation of knowledge structure is that knowledge is composed of building blocks, or key units, called schema. Schema are "interacting knowledge structures" (Rumelhart & Ortony, 1977) representing concepts stored in memory. A schema can be defined as an object "with a set of attributes that the individual perceives as being associated with the idea [object]" (Ifenthaler, Masduki, & Seel, 2011, p. 43). A learner's cognitive structure is comprised of schema and the relationships among schema (Ifenthaler et al., 2011). Cognitive structures reside in long-term memory and students must access them when confronted with a problem. Drawing upon their cognitive structure, students will produce, in working memory, a transient construct—a mental model—to solve the problem or answer a question (Ifenthaler et al., 2011; Johnson-Laird, 1983; Seel, 2003).

Students entering Introductory Biology courses likely have pre-existing schemas (albeit incomplete and possibly even incorrect) about biological concepts like "gene," "chromosome," or "fitness." These schemas represent the background knowledge onto which learners map new information. Throughout the course, students gradually expand and reorganize their cognitive structures within the domains of genetics and evolution, relating new knowledge they acquire to their pre-existing knowledge. When learning, for example, genes are located on chromosomes, an individual's *gene* and *chromosome* schemas are activated. If students' existing schemas do not account for the new information, these schemas and the relationships among them must change to accommodate the new knowledge (Rumelhart & Norman, 1978). Schema can be added to

(accretion) or slightly modified (tuning), or, when confronted with information inconsistent with any existing schema, created anew (reorganization) (Ifenthaler et al., 2011; Rumelhart & Norman, 1978). Depending on what a student's pre-existing "gene" schema is, that schema will grow and/or change as the student learns new information. Discovering that a gene's location on a chromosome has boundaries and specific landmarks may represent an example of accretion. Learning that gene expression can be turned on and off in the cell may represent an example of tuning—if it modifies a preconception of constant gene expression. Furthermore, if they believed that a gene sequence is unique and static, acquiring the concept of "allele" may lead a student to reorganize their gene schema to accommodate the idea that a gene's sequence may vary slightly by mutation and still be versions (different alleles) of the same gene. Schema are thus fluid knowledge structures; they change as students gain more knowledge about any individual schema, but more importantly, as they develop their entire cognitive structure.

Vosniadou (1994) argued that concepts are not learned in isolation but rather as part of a system. As students accumulate knowledge about individual biological schemas, they necessarily must reorganize their cognitive structure in a meaningful way to explain relationships among schemas in the context of biological systems (Ben-Zvi Assaraf & Orion, 2005; Brandstädter, Harms, & Groschedl, 2012; Hmelo-Silver et al., 2007; Jacobson & Wilensky, 2006; Liu & Hmelo-Silver, 2009; Schwarz et al., 2009). The ability to organize schemas is an important step in constructing causal models (Brewer & Nakamura, 1984; Rumelhart & Norman, 1978). Building a causal model begins with a qualitative representation of the cause and effect interactions among concepts, with focus on relationships rather than the concepts in isolation (Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Liu & Hmelo-Silver, 2009; Ruiz-Primo & Shavelson, 1996). This focus on relationships is particularly relevant as it is more representative of expert learning in science, moving from declarative knowledge, or the "what" of a system, to schematic knowledge, the "why and how" of a system (Hmelo-Silver et al., 2007; Zajchowski & Martin, 1993). A GtE mental model is composed of key genetics and evolution schemas (DNA, gene, allele, protein, phenotype, fitness) and of the functional relationships among them. Importantly, the GtE mental model is not merely the concatenation of schemas, but a representation of how these are organized to yield emergent properties, a key systems thinking tenet (Ben-Zvi Assaraf & Orion, 2005; Hmelo-Silver et al., 2007; Wilensky & Resnick, 1999). For example, knowing about genes, chromosomes, DNA, and eye color, or even the way two of these are related, does not adequately describe how genetic variation arises or how this results in phenotypic diversity or differential fitness.

When tasked with relating schema, students formulate a mental model—a representation of the relationships, processes, and strengths of interactions among schema (Johnson-Laird, 1983). An individual's schemas and their mental model feedback on one another, each possibly changing the other (Johnson-Laird, 1983). The mental model remains the ephemeral component, possibly updating the cognitive structure with new knowledge, or alternatively, the mental model may directly contradict the cognitive structure, leading to cognitive dissonance and the need for reorganization of the schema. Repeated construction of GtE mental models (during instruction and assessment) will display a students' change in conceptual understanding (Ifenthaler, 2010; Vosniadou, 1994) if the student marries new knowledge (from the new context) with their developing genetics and evolution cognitive structure. The GtE mental model is therefore the active processing unit that will change as students learn how biological systems operate and how particular concepts relate to each other. Our assumptions are that (1) cognitive structures are composed of schema and their relationships, reside in the long-term memory, and are malleable, (2) mental models develop in response to a prompt and are a product of the student's cognitive

structure, and (3) students' representations, like concept maps and models, are a reflection of their mental models (Ifenthaler et al., 2011; Novak, 1998).

Adapting Concept Maps to Model Biological Systems

Recognizing they are only partial depictions of their cognitive structures, concept maps and models serve as tools for eliciting students' mental models (Greca & Moreira, 2000; Hay, Kinchin, & Lygo-Baker, 2008; Ifenthaler, 2010; Novak, 1998; Shavelson, Ruiz-Primo, & Wiley, 2005). Quantifying the change in student-constructed models provides insight into the change in strength and clarity of connection among schema in the student's long-term memory. Indeed, research with concept maps in geology and biology courses showed major restructuring (reorganization of schema) during the first third of the course (Martin, Prosser, Trigwell, Ramsden, & Benjamin, 2000; Mintzes & Quinn, 2007; Pearsall, Skipper, & Mintzes, 1997) followed by minor restructuring (tuning) later in the course. Major restructuring was defined by changes in hierarchy as students re-conceptualized the relationships among components. More frequently, students make smaller changes to their mental models based on accretion of new knowledge and tuning of relationships (Mintzes & Quinn, 2007; Pearsall et al., 1997; Rumelhart & Norman, 1978).

Concept maps are most frequently used to organize conceptual understanding in a hierarchical manner, proceeding from most general to most specific (Novak & Canas, 2006). In an introductory course, concept maps may be particularly useful to organize declarative knowledge. We sought to focus on students' ability to organize procedural knowledge (the how and why of biological systems), rather than their entire body of knowledge (Shavelson et al., 2005). Because biological systems are complex, hierarchical, and have relationships leading to emergent properties, we applied an alternative framework more suited for representing and reasoning about systems. Goel and Stroulia (1996) proposed structure–behavior–function (SBF) theory to describe models of complex designed systems. In their framework, systems and system models are comprised of *structures* (the physical components of a system), *behaviors* (the relationships, or mechanisms connecting structures with one another), and *functions* (the roles or outputs of the system). We adapted this framework to develop a system model structured as a semantic network, where structures of a system are in boxes (nodes) and the relationships among them are described on connecting arrows (links), to illustrate how the system produces a function (Figure 1). Our goal was not to elicit the students' whole cognitive structure, but only the subset of structures and relationships relevant to explain a given function.

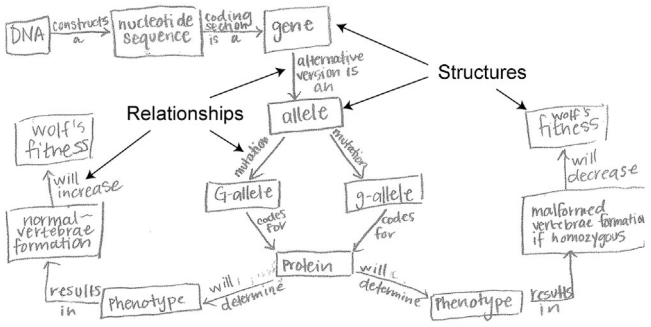

Figure 1. Student GtE model. Student-constructed model showing structures in boxes, relationships on arrows and overall representing the function of the model: show origin of genetic variation and effects of genetic variation on phenotypic variation and fitness of an organism.

We view student-constructed models as tools for facilitating students' thinking about the causal and mechanistic relationships inherent in biological systems (Long et al., in press; Vattam et al., 2011). SBF-based system models resemble concept maps but differ in two ways. First, they do not exclusively represent hierarchical relationships, but contain propositions that represent both hierarchical organization and cause–effect relationships (Sommer & Lcken, 2010). Second, SBF-based system models are not meant to represent students total domain knowledge (Jonassen, Strobel, & Gottdenker, 2005), rather, they require students to select and represent only the mechanistic relationships that are relevant to illustrate a function, or output, of a system (Hmelo-Silver & Azevedo, 2006).

In this study, we report the patterns of change observed over a semester of instruction in which, students iteratively constructed models representing their GtE mental models. We use student-generated models as a unit of analysis for quantifying and characterizing change in students' thinking about the genetic basis of evolution. Evidence of change will be measured as differences in (1) the overall architecture of students' models in terms of numbers of structures and relationships, and (2) the quality of language used in describing relationships among structures. We also investigated whether learning gains, measured in terms of model correctness, were evenly distributed across levels of achievement (GPA-based tritiles) and hypothesize that students will adapt to this tool differentially relative to their overall performance level.

Methods

Population

The study population consisted of two sections ($n = 368$) of an Introductory Biology course for life science majors at a large university in the North Central U.S. with very high research activity. We focus on two sections taught during spring semester 2009 by different primary instructors. Section 1 met three times per week for 55-minutes and was taught by two postdoctoral researchers conducting biology education research. Section 2 met twice weekly for 80 minutes and was taught by a fixed-term faculty with 5 years teaching experience at the same university. Students in the two sections did not differ demographically, and were representative of all sections of the course taught during the same semester (four sections, $n = 681$). Section demographics were representative of the university at large, but with fewer international students and a higher proportion of females in section 2 compared to the University (Table 1). Sections did not differ statistically on measures of prior achievement, including high school grade point average (GPA)

Table 1
Demographic and academic representation of students enrolled in 2 sections of Bio 1 compared to students enrolled at a large, north central USA University

		Course	
	University	Section 1 ($n = 180$)	Section 2 ($n = 188$)
Females	53%	48%	58%
Incoming GPA (max = 4)	3.4–3.8	3.1	3.2
Incoming ACT (max = 36)	23–28	25	25
International Students	11.4%	3.3%	6.4%
Freshmen		45%	47%
Sophomores		32%	32%
Life Science Majors		81%	74%

GPA and ACT values are median values. GPA, grade point average; ACT, American College Test.

and American College Test (ACT) scores, although students represented in this study achieved toward the lower ends of the ranges for students at the university. The majority of students in both sections were freshmen or sophomores and most were life science majors.

Instructional Approach

Section instructors aimed to teach as similarly as possible across both sections and met weekly to collaboratively design course materials and discuss strategies for implementation. In-class notes, activities, and assessments were virtually identical, with minor variations owing to instructor style and/or unique circumstances arising in the context of different classrooms (e.g., questions posed by students that drove particular instructional decisions).

Early in the semester (second day of class), students received instruction about model construction using SBF principles as an organizing framework (Goel & Stroulia, 1996; Hmelo, Holton, & Kolodner, 2000). In a series of brainstorming activities, students worked in cooperative groups to provide examples of scientific models, define a model, and determine commonalities among diverse examples of models discussed or displayed in class (e.g., an artist's depiction of DNA, a graph of experimental data, and a picture of the *Cnidarian* life cycle). Student feedback converged on the ideas that (a) models are "made of things" (i.e., structures), (b) the objects represented in models have relationships with other objects, and (c) models serve to simplify complex ideas or processes that cannot be directly observed, studied, or measured (i.e., functions). In this manner, students constructed an understanding of models as having structures, relationships, and functions.

Students applied their understanding of the SBF framework by constructing an initial GtE model comprised of just three structures (gene, DNA, chromosome). In-class feedback emphasized identifying relevant biological processes to explain relationships between pairs of structures in order to generate meaningful propositions. Across 7 weeks of instruction, students' models progressively incorporated additional structures and relationships, more sophisticated functions, and became contextualized for specific, real-world cases (Table 2). Students built a total of nine models either in class or as part of a homework activity during the first 7 weeks. In weeks 4, 7, and 15, students constructed a GtE model as part of a closed-book, high-stakes assessment (Table 2). From Quiz 2 to the Midterm, students added the concept of allele to their models; students in section 2 were also directed to incorporate DNA and chromosome into their Midterm GtE models. From the Midterm to the Final exam, students added the concept of fitness.

Although students were never provided with an instructor's "correct" model, extensive feedback was provided in the form of class time dedicated to comparing and evaluating peers' models, answering clicker questions or in-class problems based on patterns observed in students' models, and on-line rubrics that explained scoring as well as common issues encountered in evaluating students' models (e.g., students did not represent the origin of genetic variation effectively, arrows lacked directionality, structures were linked with incorrect biological processes, etc.).

Analyzing Model Change Over Time

Our analyses compared patterns in students' model architecture (complexity) and use of biological language (correctness) for three high stakes assessments in which students were asked to construct a GtE model. In all cases, students were provided a minimum list of structures for inclusion in students' models (Table 2). Students were encouraged to add additional structures or use structures multiple times if it helped them build a more coherent model. Students' hand-drawn models were digitized using CMAP tools (http://cmap.ihmc.us). Adjacency matrices were created for each student model on each assessment and contain the relationships between pairs of structures.

Table 2
Structures and context for each evaluated GtE model in the Introductory Biology course

	Quiz 2	Midterm	Final
Context	Not context specific	DDT resistance in mosquitoes	Deformed vertebrae in wolves
Structures[a]	**Nucleotides** **Gene** Amino acids **Protein** **Phenotype**	**Nucleotides** **Gene** Allele DNA* Chromosome* **Protein** **Phenotype**	**Nucleotides** **Gene** Allele DNA **Protein** **Phenotype** Fitness
Function	Your model should show how genes result in the production of proteins and corresponding phenotypes	Your model should show the origin of genetic variation and resulting phenotypic variation in the context of DDT resistance in mosquitoes	Your model should show 1. The origin of genetic variation among wolves; 2. The relationship between genetic variation and phenotypic variation in wolves; 3. The consequence of phenotypic variation on wolf fitness

[a]Students were allowed to add or refine structures as needed to represent the Function of the model. Bold structures are common to three assessments. Starred terms (*) were provided on the Midterm exam in section 2 but not section 1.

Complexity

We define complexity using the web-like causality index (WCI; Plate, 2010) which sums the percent of structures with greater than 1 out-arrow (a structure affects multiple structures) and the percent of structures with greater than 1 in-arrow (multiple structures affect a single structure). Although this metric was developed for causal maps, it provides a convenient objective measurement of the density of connections across models. WCI ranges from 0 (indicating a linear model) to 2 (a model in which all structures are connected with more than one other structure).

Novice mental models, as measured by a concept map, are perceived to be less integrated and less complex compared to experts (Ifenthaler et al., 2011). We do not necessarily assume the same for expert versus novice performance when measured by system models. SBF theory suggests the external representation is meant to convey the function of the system, not an individual's complete domain knowledge. As such, it would be reasonable to predict that experts would build models that more accurately and efficiently communicate system function (i.e., more parsimonious) compared to novices. For the purposes of this research, we do not have *a priori* hypotheses about the direction of changes in complexity through time; rather, we are interested in whether change in complexity in student-constructed models can provide insight into changes in students' cognitive structures during the learning process.

Correctness

We define correctness as the mean biological accuracy of all the relationships in a student's model. Using thousands of student responses, we constructed a grounded correctness rubric for all

Table 3
Rubric for rating relationships in student-constructed GtE models

Code	Interpretation	Example
1. Incorrect	Student's use of language does not explain the relationship between a pair of structures. Or, student established a link between structures by drawing an arrow, but failed to include words describing the relationship.	Chromosome → constructed by → DNA
2. Marginal	Student conveys some understanding of the relationship between structures, but their specific use of language is somewhat vague, falling short of an ideal, or biologically accurate, explanation.	Chromosome → contains → DNA
3. Accurate	Student's use of language is biologically accurate and reflects mastery understanding for college-level Introductory Biology.	Chromosome → made of two strands of → DNA

possible propositions ([structure] → relationship → [structure] combinations). Scoring propositions (e.g., [chromosome] → coding section is → [gene], see Table 1 in Supporting Information Appendix for more examples) is a valid measure of students procedural and structural systems thinking (Brandstädter et al., 2012). Relationships were rated on a scale of 1–3 indicating the biological accuracy of students' language (Table 3).Four raters achieved 77% agreement using the rubric with 10 models (total of 112 propositions). The rubric was adjusted to resolve discrepancies among raters. Additional relationships were added to the rubric over time as they were encountered in student models and the rating for each new behavior was verified among raters. The rubric is provided as Supporting Information Appendix (Table 1).

Statistical Analyses

Haak et al. (2011) showed that incoming GPA can be a good predictor of student performance and showed that active learning practices may impact students differently depending on their prior academic achievement. Therefore, we compared outcomes for students in different achievement tritiles based on students' cumulative university GPA upon entering the course and evenly partitioned as: Upper tritile = GPAs > 3.46; Middle tritile = GPA between 2.89 and 3.46; Lower tritile = GPA < 2.89. A linear mixed effects model with repeated measures was constructed to determine the independent and interacting effects of section, assessment, tritile, and model complexity (WCI) on students' mean model correctness. Analyses were conducted in R using the nlme library (Pinheiro, Bates, DebRoy, Sarkar, & the R Development Core Team (2-01), 2012).

Results

Complexity

Student-constructed models changed in both complexity and overall correctness throughout the course. Complexity increased from Quiz 2 to the Midterm, but then decreased from Midterm

to Final exam. Model architecture is a function of the number of independent structures and number of relationships connecting structures. Models constructed on Quiz 2 stand out for their structural differences from other assessments. Nearly all students (89%) constructed linear models (WCI = 0) using five structures and four relationships (Figure 2 and Table 4). Following Quiz 2, students began using a greater number of relationships than structures and the number of linear models decreased to 4% and 5% on the Midterm and Final, respectively (Figure 3).

At the Midterm, students in section 1 used twice as many structures as provided despite being provided fewer structures in the item instructions compared to section 2 (Tables 2 and 4). Section 2 students used proportionally more relationships relative to structures and had a higher WCI (0.42) than section 1 students (0.35). Overall, Midterm model WCI scores were 0.38 (SE = 0.012), revealing that on average, students connected 19% (0.38 divided by maximum WCI score of 2) of structures to more than 1 other structure (greater connectivity). On the other hand, more than 80% of the structures were only connected to one other structure.

On the Final exam, students in both sections were provided seven structures and the number of structures and relationships only increased marginally (Table 4). Compared with the Midterm, section 1 students added 0.6 structures on the Final despite being provided with 2 additional structures compared to the Midterm. Section 2 students were provided with the same number of structures on the Midterm and Final and yet they added 1 additional structure to their Final exam model. There was a parallel increase in number of relationships in both sections. Across both sections, the WCI decreased from 0.38 to 0.34 (SE = 0.012) meaning the number of structures connected to only one other structure rose to more than 85% by the Final exam.

Model Correctness and Interactions Among Factors

More than 10,000 relationships across all three models in both sections were rated for their biological accuracy with 27% rated as a level 1 (incorrect), 44% rated as a level 2 (correct but vague), and 28% rated as a level 3 (best we would expect from an Introductory Biology student). Mean model correctness was not normally distributed (Shapiro–Wilks, $W = 0.982$, $p < 0.001$), although the quantile–quantile plot approximated a normal relationship, the data were mesokurtic, and the data were not skewed. The number of relationships in a model could potentially affect both

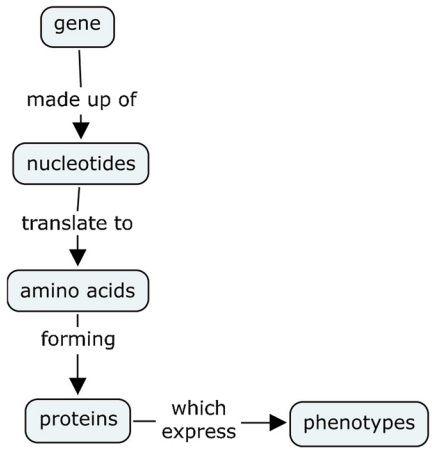

Figure 2. Average Quiz 2 model. On Quiz 2, students created very linear models with relationships usually rated below 2 (marginal). This model represents the mean complexity and mean correctness for Quiz 2 models.

Table 4
Mean (SE) number of structures and relationships used in student-constructed models on three assessments

	Overall	Section	
		1	2
Quiz 2			
Structures	4.9 (0.05)	4.8 (0.07)	4.9 (0.06)
Relationships	3.9 (0.04)	3.9 (0.06)	4.0 (0.06)
Midterm			
Structures[a]	10.1 (0.17)	10.6 (0.23)	9.6 (0.26)
Relationships	11.0 (0.23)	11.3 (0.28)	10.8 (0.36)
Final			
Structures	10.9 (0.17)	11.2 (0.19)	10.5 (0.29)
Relationships	11.6 (0.22)	11.7 (0.24)	11.4 (0.37)

[a]Students in section 1 were presented with 5, 5, and 7 structures while section 2 was presented with 5, 7, and 7 structures on successive assessments. See Table 2 for structures provided.

mean correctness and WCI. However, mean model correctness was found to be independent of the number of structures and relationships included in an individual's model for the Midterm and Final exams ($F = 0.19$, $p = 0.66$, $R^2 < 0.01$; $F = 3.30$, $p = 0.07$, $R^2 = 0.01$, respectively). There was insufficient variance in number of structures and relationships on Quiz 2 to include in the analysis. A mixed-effects fitted regression model identified section, tritile, and assessment

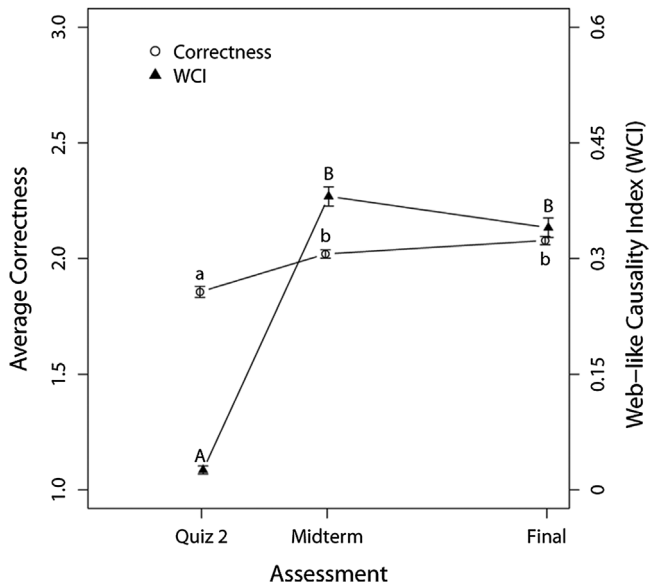

Figure 3. Model correctness and complexity. Mean correctness of relationships used in student models increased throughout the semester. The web-like causality index (WCI) increased in the first half of the semester and these gains were retained through the end of the course. Values denoted with similar letters are not significantly different. Capital letters are comparing WCI values and lower-case letters are comparing mean correctness scores.

having significantly affected students' mean model correctness (all $p < 0.05$). WCI was not a significant factor in students' mean correctness score ($F = 2.12, p = 0.15$).

Students in section 1 created models with significantly greater mean correctness scores (2.02, SE = 0.016) than students in section 2 (1.94, SE = 0.017), a pattern that was evident in each assessment. Across sections, mean model correctness increased significantly from Quiz 2 (1.86, SE = 0.02) to the Midterm exam (2.02, SE = 0.02, $z = 7.44$, $p < 0.001$) but increased only marginally to the Final exam (2.08, SE = 0.02, $z = 1.6, p = 0.25$, Figures 2 and 4).

Overall, upper tritile students used significantly more ($F = 41.13$, $p < 0.001$) correct relationships (2.12, SE = 0.018) than students in the middle tritile (2.01, SE = 0.021) and lower tritile (1.83, SE = 0.021). Tritiles differed significantly in model correctness, but all improved on successive assessments with no significant interaction of tritile and assessment ($p = 0.33$; Figure 5). Mean correctness for lower tritile students remained below the threshold correctness level of 2 for all assessments, although they improved from 51% to 67% of their relationships rated >2 by the Final exam. Middle tritile students had a mean correctness below 2 on their Quiz 2 models, then increased their mean correctness to 2.03 and 2.10 on their Midterm and Final models, respectively. Conversely, mean correctness for Upper tritile students remained at or above the threshold correctness level of 2 throughout the course. For Upper tritile students at the Final exam, more than 35% of their relationships were rated as a 3 and 46% rated as level 2.

Interestingly, the correctness gap among tritiles' narrowed during the course. On Quiz 2, the mean correctness gaps separating the Lower and Middle tritiles from the Upper tritile were 18% and 12%, respectively (Figure 5). These gaps narrowed to 15% and 5% by the Final exam, suggesting that although both the Lower and Middle tritiles were reducing the achievement gap separating them from the highest performing students, Middle tritile students were doing so more quickly.

Discussion

In this study, we used iterative model construction as a way to explore change in students' mental models during a semester of instruction. We observed that students constructed increasingly correct GtE models with concomitant change in model architecture. From Quiz 2 to the Midterm, students' models progressed from predominantly linear structures to models with higher complexity in the form of branching and interconnectedness among structures. From the Midterm to Final, model complexity declined while correctness continued to increase, suggesting

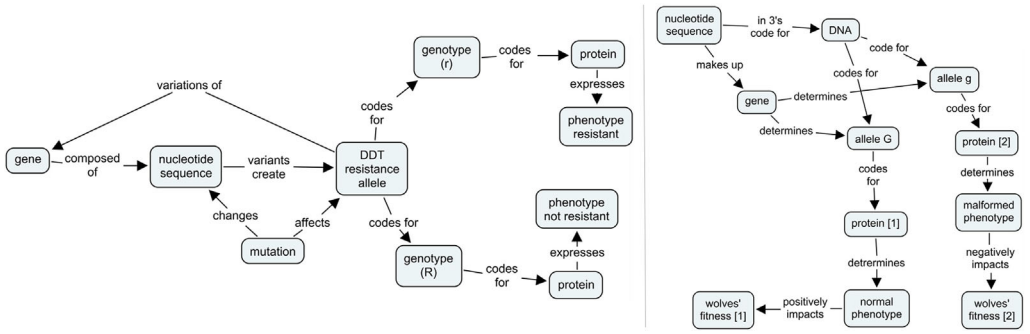

Figure 4. Average Midterm and Final models. Students increased their correctness on their Midterm and Final exam models. Complexity increased from Quiz 2 to Midterm exam, then decreased slightly to the final. Selected models represent the mean complexity and mean correctness for the respective exams.

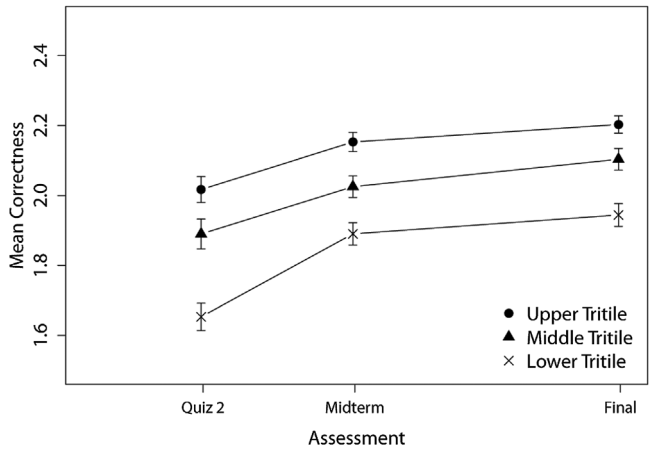

Figure 5. Correctness gap. Mean correctness of relationships in student-constructed models increases for all tritiles from Quiz 2 to the Midterm exam, then increased marginally to the Final exam. The gap in correctness between the Upper tritile and the other two tritiles decreased during the course.

that students were building more parsimonious models by the end of the course. We hypothesize that the observed changes in model construction can be explained by students' accretion, tuning, and reorganization of schemas in order to explain the genetic foundations of evolution. Students' schemas were consistently activated during the course in a manner that emphasized relationships among them. Early in the course, students were accreting knowledge about biological concepts and learning how to appropriate that knowledge in the context of a modeling task. In the later half of the semester. This required students to frequently update and refine their schema as they learned new content and, most importantly, how schemas were related in novel scenarios.

Accretion

Developing one's schemas begins with accumulation of knowledge about concepts (i.e., improving their definition) and learning how to relate pairs of concepts to each other (i.e., expanding the definition to include other concepts). Pearsall et al. (1997) showed a tripling in the number of concepts and relationships in students' concept maps early in a semester attributable to accretion and restructuring of knowledge. We similarly observed an increase in the number of concepts and relationships in students' models early in the semester that could be explained by their accretion of new knowledge. Early-semester formative assessments (e.g., clicker questions, pre-tests) indicated that many students could recall definitions of genetics concepts (e.g., gene, chromosome). However, Quiz 2 revealed that most students produced linear models with low biological accuracy. We suspect that despite knowing definitions of concepts, students had little or no practice *relating* biological concepts to each other prior to the course. Specifically, they had not been asked to describe *how* genes affect phenotypes.

Biologists are interested in more than definitions and seek to understand concepts through their interactions with other concepts. During the early period in the course, students are likely accreting knowledge about biological relationships as well as the concepts themselves. Since many of these relationships did not exist previously in students' cognitive structures, students are challenged to accommodate the new knowledge and to communicate it in the form of external representations. Differences among tritiles on Quiz 2 may be explained by differences in magnitude, rates, or timing of accretion. Upper tritile students may have had to accrete less than

other students if they entered the course with greater knowledge about the biological concepts and/or their relationships to one another. Hence, Lower and Middle tritile students would struggle more to form a cohesive mental model if they were concurrently accreting knowledge about concepts (e.g., alleles, genes), as well as their relationships to other concepts across microscopic and macroscopic scales.

Major Restructuring

In our modeling task, students cannot explain model function by simply expanding or connecting their schemas. Rather, students must organize their schemas in a meaningful way that is more representative of how biologists use models to represent or reason about biological phenomena. Mintzes and Quinn (2007) defined major restructuring in concept maps as the "introduction of powerful new organizing concepts that subsume existing ideas and forge fundamentally novel explanatory or descriptive frameworks of knowledge" (p. 283). They quantify major restructuring by increased frequency of hierarchies, that is, organizing an increased number of concepts under more general concepts. We conceive of major restructuring in SBF-based system models when students concurrently increase complexity (greater interconnectivity) and maintain or increase the biological correctness of relationships (Figure 6). Major restructuring would mean these students have recognized and can depict a key system model component: multiple cause and effect. We observe this in student models when they show how mutation in the genome results in multiple alleles yielding multiple phenotypes. Student-constructed models must necessarily incorporate branching and additional structures and relationships in order to capture this process. Despite a prevalence of linear models on Quiz 2, more than 95% of students created a nonlinear (more interconnected) model on the Midterm exam. More than half the students (59%) increased complexity and correctness between Quiz 2 and the Midterm. Did all of

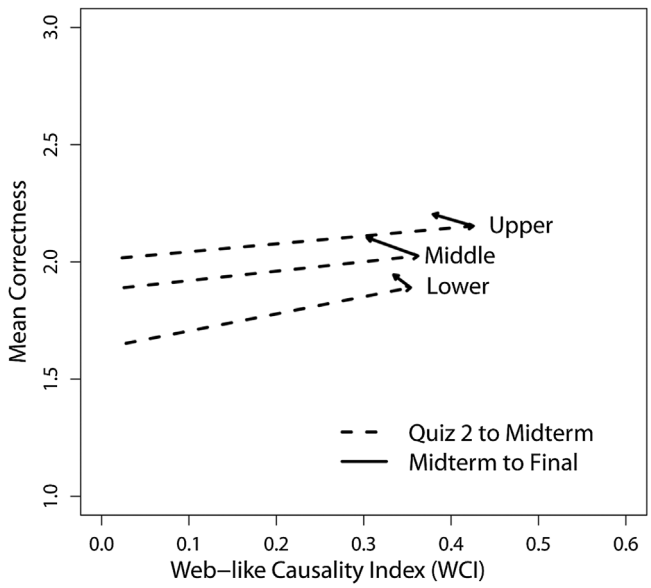

Figure 6. Change in correctness and complexity. All tritiles expanded and improved the biological correctness of their models from Quiz 2 to the Midterm. Later in the course, the student-constructed models continued to increase in biological correctness, but were pruned to include fewer connecting relationships.

these students undergo major restructuring between Quiz 2 and the Midterm even though this requires substantial mental effort (Rumelhart & Norman, 1978)?

On Quiz 2, approximately 8% of students constructed branched models. The same students also created more biologically correct models than their tritile peers on Quiz 2. For these students, we cannot determine whether they experienced restructuring prior to Quiz 2 or whether they entered the course with more appropriate mental models. Not surprisingly, these students only increased marginally in complexity and correctness from Quiz 2 to the Midterm.

Our data suggest that most students restructured their knowledge between Quiz 2 and Midterm, and may have been in response to classroom feedback and interactions with peers following Quiz 2. Both sections provided rubrics with qualitative criteria for grading (e.g., 1 pt for each link, 1 pt for providing an appropriate behavior). The rubric for section 1 also provided 3 examples of acceptable student-derived models (including both linear and branched models), but did not provide an "expert" model or list of "correct" relationships. Section 2 opted not to include any exemplar models on the rubric. Both sections provided in-class feedback 1 week following Quiz 2, but differed somewhat in their approach. Students in section 1 completed a modeling activity where students suggested what structures and relationships to incorporate in a model, while the instructor recorded students' additions on the whiteboard. As a result, students developed a consensus model as a class, incorporating multiple possible relationships, and explicitly discussed choice and placement of structures and relationships. Students in section 1 were, therefore, afforded an opportunity to evaluate their own models in relation to a consensus model (Kinchin, Hay, & Adams, 2000). Students in section 2 worked in small groups to complete a worksheet activity in which they evaluated 3 peer models that appeared on Quiz 2 that varied in their degree of branching, correctness of relationships, and appropriateness in describing the assigned function. Worksheet questions asked students to reflect on each model's effectiveness in communicating the function and the role of branching in model construction. Groups then constructed a consensus model based on their discussions. Restructuring may have been prompted for students in both sections because they were afforded an opportunity to compare and evaluate models in a structured format (Gentner & Markman, 1997; Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000). The intention of feedback was to enable metacognitive thinking about students' own modeling and to highlight strengths and weaknesses of alternative approaches to modeling a function (Van Meter, 2001). The change in complexity of models following Quiz 2 may reflect students' learning about the advantage of showing multiple relationships and/or using branching as a tool for illustrating variation. Alternatively, students may increase complexity if they perceive instructor bias towards a more branched model typology.

The Midterm provides a snapshot suggesting that most students restructured following Quiz 2, but differ in their specific pathway. Few students constructed a linear model on both Quiz 2 and the Midterm, but among those who did, most increased their correctness on the Midterm. These students appear to be improving their conceptual correctness prior to incorporating multiple relationships in their models. Conversely, a larger number of students (31%, $n = 109$) appear to be restructuring by first accommodating variation in their model (i.e., utilizing branching to show the origin of multiple alleles), while the biological correctness of their relationships declines. One might expect biological correctness to decline as mental effort is allocated to ordering concepts and showing a particular function. Once concepts have been organized, we could reasonably expect to see effort once again placed on improving the biological correctness of relationships. Most of these students (78 of the 109) increased their correctness (by 0.21) by the Final exam, suggesting that the difficult reorganization observed at the Midterm was ultimately productive in their ability to construct an acceptable model. Additional summative assessments could provide clarity about specific patterns in students' thinking during this period of rapid conceptual change.

Minor Restructuring (Tuning)

After the major mental restructuring earlier in the course, students were primed to integrate more knowledge into their developing GtE mental model and focus on the relationships among the schemas. Mintzes and Quinn (2007) and Pearsall et al. (1997) define weak restructuring or tuning of concept maps by increases in structures and relationships (using SBF language) without increases in branches and cross-links. We believe weak restructuring in GtE models occurs when students continue to improve the biological correctness of their models while increasing model parsimony, pruning relationships that do not contribute explicitly to the function of the model. From the Midterm to the Final exam, students removed connections (decreased complexity) that were inappropriate or poorly understood while focusing on improving the quality of connections (correctness) that were most important to the model function. At the time of the Midterm exam, students had just begun to accrete knowledge about how phenotypic variation affects organismal fitness—a major focus of the third quarter of the course. It is possible that the genetic basis for phenotypic variation had already been solidified through scaffolded modeling, which enabled students to add fitness (simple cause and effect relationship from phenotype) with relative ease.

We acknowledge the potential of a ceiling for complexity and biological correctness that is reflective of the research design rather than student ability. Model complexity would surely increase if students were asked to include more than two alleles, multiple interacting genes, or fitness effects on allele frequency. Although we did not explicitly test this, we predict that the modeling skills students developed during the course would allow students to tune their mental models in order to incorporate additional structures or functions. Complexity may also have been constrained by the instructional goal of the GtE model itself. Instructors sought to use the GtE model as a way for students to connect molecular-level processes to population-level outcomes in a way that was broadly transferable across a range of biological cases and contexts. A completely interconnected model (every structure relating to every other structure) yields a WCI of 2. Although many of the structures provided to students could have been interconnected, connecting them would have resulted in awkward or meaningless relationships. For example, [genes] → increase → [fitness] may be partially correct, but ignores the myriad functions that cause the increase. These connections are possible, but not technically correct. GtE assessment items were constructed to elicit a representation depicting concepts that are strongly connected to describe a function. Students were not asked to show their complete domain knowledge, as might be expected if using concept maps of causal systems (Stewart, 2012).

The correctness scale ranged from 1 to 3, but level 2, indicating acceptable but not ideal responses, was disproportionately represented in ratings of student responses. An expansive median level (44% of relationships) led to a conservative estimate of students' mean correctness. The potential correctness constraint is also evident in the gap between tritiles. Lower and Middle tritiles closed the gap with the Upper tritile either because they improved more quickly or because the Upper tritile reached a plateau earlier (Figure 5). One potential explanation is that models help Lower and Middle tritile students to a greater extent than students who are already high-performing. Haak et al. (2011) found active learning methods disproportionately help underrepresented minorities. Model-construction is an example of an active learning method that may produce a similar effect. By focusing students' attention less on definitions and more on integrating new information and connecting existing ideas, models may help struggling students better organize their knowledge and generate "big picture" understandings. Interestingly, no Upper tritile students used all level 3 relationships on their Final exam models and more than half (65 out of 116) had more than one incorrect relationship. While the Upper tritile contains excellent

students, half continued to use vague language for 2 or more relationships. Although Upper tritile students did not reach level 3 understanding of all relationships, they still demonstrated greater understanding than their peers. Perhaps future studies should include GtE model construction in upper division genetics or evolution courses in studies similar to Mintzes and Quinn (2007) to quantify the longitudinal effect of tritile differences and determine if a ceiling exists on students' ability to relate these concepts.

Implications for Instruction and Assessment

Students entering Introductory Biology often have a decade or more of education requiring knowledge of isolated concepts, yet biologists routinely consider concepts in relation to each other. Documents such as Next Generation Science Standards (National Research Council, 2012) and Vision and Change (AAAS, 2009), indicate momentum for change across the curriculum that emphasizes *connections* among concepts and enhancing systems thinking and modeling skills critical to 21st century learners. As students enter college-level biology, instructors should assess students' modeling skills and provide opportunities to improve on those skills in order to facilitate learning across a broad suite of biological systems, including food webs (community dynamics), coupled biogeochemical cycles (cycling of matter and energy), and central dogma.

Moving from content learning to learning in the context of systems is neither simple nor fast and instructors must consider how to best facilitate this transformation in their courses. Understanding concepts in terms of their relationships with other concepts represents a fundamentally different way of thinking for many students. Our work shows this conversion takes time and students proceed at different speeds. As a starting place, we advocate integrating modeling early in biology instruction by asking students to model relationships among a small number of concepts. One might perceive this task as remedial, but the complexity and correctness of Quiz 2 models shows that students do not find the task remedial. As with other skills, such as quantitative reasoning and argumentation(Hung, 2008; Jimenez-Aleixandre, Rodriguez, & Duschl, 2000), modeling requires practice (Brewe, 2008) and students will benefit from multiple opportunities to apply and revise their mental models in new contexts. Iterative model construction helps students become metacognitive about their modeling ability and consider models as changeable or improvable entities that are generalizable to multiple phenomena (Schwarz et al., 2009; Van Meter, 2001).

Integrating modeling as a core component of pedagogy represents a commitment where instructors attend to students' progression and tailor feedback to meet students' specific needs. Modeling provides the instructor with critical feedback about what students are learning, as well as the rate at which they are incorporating new knowledge (Hay et al., 2008). In our experience, data derived from student-constructed models have proved informative about students' content learning as well as their modeling skills. Directed feedback about both will prepare students for adding conceptual complexity to their models, adapting models for different contexts and cases, and strengthening understanding about conceptual relationships.

Constructing biology models may be a step towards improving systems thinking, but additional research is needed to affirm such outcomes and elucidate underlying mechanisms. Model construction may improve systems thinking by focusing students less on specific content and more on conceptual relationships, their use of language, overall model functionality, and system phenomena (Booth Sweeney & Sterman, 2007; Liu & Hmelo-Silver, 2009). Our results indicate students improved their structural system thinking, that is, identifying the concepts and relationships within a system (Brandstädter et al., 2012). However, we did not test whether students were also advancing towards procedural system thinking by understanding the dynamic and emergent properties inherent in a biological system (Brandstädter et al., 2012).

To adopt modeling as a practice and not a task, "students need an authentic reason for building a model other than 'doing school"' (Schwarz et al., 2009, p. 652). In our design of Introductory Biology, modeling served as a tool for strengthening and connecting students' understanding of molecular-level phenomena in order to explain evolutionary outcomes. Whether reasoning about a problem in human medicine, agriculture, or conservation, students will need to apply the same explanatory framework to describe how genetic variation arises, is expressed by organisms, and can be acted on by the environment. Our goal was to help students develop and refine a foundational cognitive structure they could apply broadly throughout the domain of biology and perhaps inspire the use of modeling in non-classroom contexts. Biology educators are largely practicing biologists that regularly and intuitively use models to reason through problems and explain their thinking to others. As a community, perhaps one of the simplest things we can do to better prepare future biologists is to draw upon the practices and ways of thinking that are second nature for us. Integrating modeling throughout a biology curriculum has the potential to transform our students from linear, cause-effect thinkers to individuals who are able to reason about real-world problems that are complex, multi-scale, and have multiple causes and effects.

We thank D. Ebert-May and S. Wyse for their collaboration in the design of assessments and rubrics and for their feedback regarding implementation of models in their courses. We thank the editors and reviewers for helpful comments on earlier versions of this article. We thank K. Kostelnik for assistance in improving the correctness rubric and A. George, M. Gustafson, J. LaCrosse, D. Forney, and A. Makohon-Moore for logistical and technical support throughout the project. This material is based upon work supported by the National Science Foundation under Grant No. DRL 0910278. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the National Science Foundation.

References

AAAS. (2009). Vision and change in undergraduate biology education: A call to action. Washington, DC: American Association for the Advancement of Science.

Bahar, M., Johnstone, A. H., & Hansell, M. H. (1999). Revisiting learning difficulties in biology. Journal of Biological Education, 33(2), 84–86. doi: 10.1080/00219266.1999.9655648

Ben-Zvi Assaraf, O., & Orion, N. (2005). Development of system thinking skills in the context of earth system education. Journal of Research in Science Teaching, 42(5), 518–560.

Booth Sweeney, L., & Sterman, J. D. (2007). Thinking about systems: Student and teacher conceptions of natural and social systems. System Dynamics Review, 23(2–3), 285–311.

Brandstädter, K., Harms, U., & Großschedl, J. (2012). Assessing system thinking through different concept-mapping practices. International Journal of Science Education, 34(14), 2147–2170. doi: 10.1080/09500693.2012.716549

Bray-Speth, E., Long, T. M., Pennock, R. T., & Ebert-May, D. (2009). Using Avida-ED for teaching and learning about evolution in undergraduate Introductory Biology courses. Evolution: Education and Outreach, 2, 415–428.

Brewe, E. (2008). Modeling theory applied: Modeling instruction in introductory physics. American Journal of Physics, 76, 1155.

Brewer, W. F., & Nakamura, G. V., (1984). *The nature and functions of schemas* (No. 325). University of Illinois at Urbana-Champaign. Retrieved from https://www.ideals.illinois.edu/handle/2142/17542

Feltovich, P. J., Coulson, R. L., & Spiro, R. J. (2001). Learners' (mis)understanding of important and difficult concepts: A challenge to smart machines in education. In K. Forbus & P. J. Feltovich (Eds.), Smart machines in education: The coming revolution in educational technology (pp. 349–375). Menlo Park, CA: AAAI/MIT Press. Retrieved from http://dl.acm.org/citation.cfm?id=570962

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. American Psychologist, 52(1), 45–56. doi: 10.1037/0003-066X.52.1.45

Goel, A., & Stroulia, E. (1996). Functional device models and model-based diagnosis in adaptive design. Artificial Intelligence for Engineering, Design, Analysis and Manufacturing, 10(04), 355–370. doi: 10.1017/S0890060400001670

Greca, I. M., & Moreira, M. A. (2000). Mental models, conceptual models, and modelling. International Journal of Science Education, 22(1), 1–11. doi: 10.1080/095006900289976

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in Introductory Biology. Science, 332(6034), 1213–1216. doi: 10.1126/science.1204820

Hay, D., Kinchin, I., & Lygo-Baker, S. (2008). Making learning visible: The role of concept mapping in higher education. Studies in Higher Education, 33(3), 295–311.

Hmelo, C. E., Holton, D. L., & Kolodner, J. L. (2000). Designing to learn about complex systems. Journal of Learning Science, 9(3), 247–298.

Hmelo-Silver, C. E., & Azevedo, R. (2006). Understanding complex systems: Some core challenges. Journal of the Learning Sciences, 15(1), 53–61.

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert-novice understanding of complex systems. Journal of Learning Science, 16(3), 307–331.

Hung, W. (2008). Enhancing systems-thinking skills with modelling. British Journal of Educational Technology, 39(6), 1099–1120.

Ifenthaler, D. (2010). Relational, structural, and semantic analysis of graphical representations and concept maps. Educational Technology Research and Development, 58(1), 81–97.

Ifenthaler, D., Masduki, I., & Seel, N. M. (2011). The mystery of cognitive structure and how we can detect it: Tracking the development of cognitive structures over time. Instructional Science, 39(1), 41–61.

Jacobson, M. J., & Wilensky, U. (2006). Complex systems in education: Scientific and educational importance and implications for the learning sciences. Journal of the Learning Sciences, 15(1), 11–34.

Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "doing science:" Argument in high school genetics. Science Education, 84(6), 757–792.

Johnson-Laird, P. N. (1983). Mental models: Towards a cognitive science of language, inference, and consciousness. Cambridge, MA: Harvard University Press.

Johnstone, A. H. (1991). Why is science difficult to learn? Things are seldom what they seem. Journal of Computer Assisted Learning, 7(2), 75–83.

Jonassen, D., Strobel, J., & Gottdenker, J. (2005). Model building for conceptual change. Interactive Learning Environment, 13(1–2), 15–37.

Kinchin, I. M., Hay, D. B., & Adams, A. (2000). How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. Educational Research, 42(1), 43–57.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), Causal learning: Psychology, philosophy, and computation (pp. 154–172). Oxford, UK: Oxford University Press.

Liu, L., & Hmelo-Silver, C. (2009). Promoting complex systems learning through the use of conceptual representations in hypermedia. Journal of Research in Science Teaching, 46(9), 1023–1040.

Long, T., Dauer, J., Kostelnik, K., Momsen, J., Wyse, S., Bray Speth, E., & Ebert-May, D. Fostering ecoliteracy through model-based instruction. Frontiers in Ecology and the Environment (in press).

Marbach-Ad, G., & Stavy, R. (2000). Students' cellular and molecular explanations of genetic phenomena. Journal of Biological Education, 34(4), 200–205. doi: 10.1080/00219266.2000.9655718

Martin, E., Prosser, M., Trigwell, K., Ramsden, P., & Benjamin, J. (2000). What university teachers teach and how they teach it. Instructional Science, 28, 387–412.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The Influence of Shared Mental Models on Team Process and Performance. Journal of Applied Psychology, 85(2), 273–283. doi: 10.1037/0021-9010.85.2.273

Mervis, J. (2011). Weed-out courses hamper diversity. Science, 334(6061), 1333. doi: 10.1126/science.334.6061.1333

Mintzes, J., & Quinn, H. J. (2007). Knowledge restructuring in biology: Testing a punctuated model of conceptual change. International Journal of Science and Mathematics Education, 5, 281–2306.

National Research Council, (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. committee on a conceptual framework for new K-12 science education standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nehm, R., Poole, T. M., Lyford, M. E., Hoskins, S. G., Carruth, L., Ewers, B. E., & Colberg, P. J. (2009). Does the segregation of evolution in biology textbooks and introductory courses reinforce students' faulty mental models of biology and evolution? Evolution: Education and Outreach, 2(3), 527–532.

Nehm, R., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. BioScience, 57(3), 263–272.

Nehm, R., & Ridgway, J. (2011). What do experts and novices "see" in evolutionary problems? Evolution: Education and Outreach, 4(4), 666–679.

Nehm, R., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. Journal of Research in Science Teaching, 45(10), 1131–1160.

Novak, J. D. (1998). Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations. Mahwah, NJ: L. Erlbaum Associates.

Novak, J. D., & Canas, A. J. (2006). The origins of the concept mapping tool and the continuing evolution of the tool. Information Visualization Journal, 5(3), 175–184.

Pearsall, N. R., Skipper, J. E. J., & Mintzes, J. J. (1997). Knowledge restructuring in the life sciences: A longitudinal study of conceptual change in biology. Science Education, 81(2), 193–215.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & the R Development Core Team (2-L 01). (2012). nlme: Linear and nonlinear mixed effects models.

Plate, R. (2010). Assessing individuals' understanding of nonlinear causal structures in complex systems. System Dynamics Review, 26(1), 19–33. doi: 10.1002/sdr.432

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. Journal of Research in Science Teaching, 33(6), 569–600.

Rumelhart, D. E., & Norman, D. (1978). Accretion, tuning and restructuring. In J. Cotton & R. Klatzky (Eds.), Semantic factors in cognition (pp. 37–54). Hillsdale, NJ.: Erlbaum Lawerence Associates.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson R. J. Spiro & W. E. Montague (Eds.), Schooling and the Acquisition of Knowledge. (pp. 99–135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D. . . . Krajcik, J., (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. Journal of Research in Science Teaching, 46(6), 632–654.

Seel, N. (2003). Model-centered learning and instruction. Technology, Instruction, Cognition and Learning, 1(1), 59–85.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (2005). Windows into the mind. Higher Education, 49(4), 413–430.

Sinatra, G. M., Brem, S. K., & Evans, E. M. (2008). Changing minds? Implications of conceptual change for teaching and learning about biological evolution. Evolution: Education and Outreach, 1(2), 189–195.

Smith, M. U. (2010a). Current status of research in teaching and learning evolution: II. Pedagogical issues. Science & Education, 19(6–8), 539–571.

Smith, M. U. (2010b). Current status of research in teaching and learning evolution: I. Philosophical/epistemological issues. Science & Education, 19(6–8), 523–538.

Sommer, C., & Lücken, M. (2010). System competence—Are elementary students able to deal with a biological system? Nordic Studies in Science Education, 6(2), 125–143.

Stewart, M. (2012). Joined up thinking? Evaluating the use of concept-mapping to develop complex system learning. Assessment & Evaluation in Higher Education, 37(3), 349–368.

Van Meter, P. (2001). Drawing construction as a strategy for learning from text. Journal of Educational Psychology, 93(1), 129.

Vattam, S. S., Goel, A. K., Rugaber, S., Hmelo-Silver, C. E., Jordan, R., Gray, S., & Sinha, S. (2011). Understanding complex natural systems by articulating structure–behavior–function models. Journal of Educational Technology and Society, 14(1), 66–681.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. Learning and Instruction, 4(1), 45–69. doi: 10.1016/0959-4752(94)90018-3

Wilensky, U., & Resnick, M. (1999). Thinking in levels: A dynamic systems approach to making sense of the world. Journal of Science Education and Technology, 8(1), 3–19.

Zajchowski, R., & Martin, J. (1993). Differences in the problem solving of stronger and weaker novices in physics: Knowledge, strategies, or knowledge structure? Journal of Research in Science Teaching, 30(5), 459–470. doi: 10.1002/tea.3660300505

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.

Supporting Information Table 1. Correctness Rubric.