


RESEARCH ARTICLE

Open Access



Iterative design of a simulation-based module for teaching evolution by natural selection

Jody Clarke-Midura^{1*} , Denise S. Pope², Susan Maruca³, Joel K. Abraham⁴ and Eli Meir³

Abstract

Background: This research builds on a previous study that looked at the effectiveness of a simulation-based module for teaching students about the process of evolution by natural selection. While the previous study showed that the module was successful in teaching how natural selection works, the research uncovered some weaknesses in the design. In this paper, we used design-based research to investigate how design changes to the module affected not only students' understanding of the concepts but also their usage of misconceptions in the assessments. We present results from two studies. In study 1, we looked at gains in understanding on a pre and post-assessment for students who used the revised version of the module. We also examined misconception uses in their answer selections. In study 2, we compared the performance on a summative assessment between students who used the revised version and students who used the original version of the module. We also looked at misconception uses in their answer selections.

Results: In study 1, we saw a significant improvement in the pre-post assessment for students who used the revised version. In study 2, we did not find a significant difference on the overall performance outcome between students who used the revised and those that used the original version of the module. In both studies, however, we saw a lower use of misconceptions after students used the revised module. In particular, we saw less use of the adaptive mutation misconception, the belief that mutations are adaptive responses to the environment and are biased towards advantageous mutations. This is promising because in the previous study there was no evidence of decreased use of this misconception.

Conclusions: Students showed learning gains on all targeted key concepts, and reduced expression of all targeted misconceptions, which was not found previously for students using the older workbook version of the module. In particular, the revised version appears to help students overcome the adaptive mutation misconception. This article demonstrates how design-based research can contribute to the ongoing improvement of evidence-based instruction in undergraduate biology classrooms.

Keywords: Simulation, Natural selection, Evolution, Misconceptions, Undergraduate

Background

It is well-documented that the theory of evolution by natural selection is difficult for many students to learn (reviewed in Gregory 2009; Bishop and Anderson 1990), with many misconceptions persisting after instruction

(Bishop and Anderson 1990; Ferrari and Chi 1998; Kalinowski et al. 2013). This persistence supports the need for a variety of instructional approaches to teach the process (Nehm and Reilly 2007), and the value of returning to the topic multiple times across an individual course and in multiple courses, at all academic levels (Kalinowski et al. 2013).

Active learning approaches to teaching complex topics such as evolutionary theory are demonstrably more

*Correspondence: jody.clarke@usu.edu

¹ Department of Instructional Technology and Learning Sciences, Utah State University, 2830 Old Main Hill, Logan, UT 84321, USA
Full list of author information is available at the end of the article

effective than passive approaches such as lectures and textbook readings (Freeman et al. 2014). Kalinowski et al. (2013) suggest there is a shortage of instructional exercises that have been designed specifically to address natural selection misconceptions and that require active participation of students, and many techniques and tools used to teach evolution by natural selection have not been subject to rigorous assessment of their utility (Nehm 2006). One approach for developing and assessing instructional tools is design-based research (DBR), a methodological approach that attempts to better understand students' learning through a continuous iteration of design, implementation, analysis, and redesign of an intervention in an authentic context (DBR Collective 2003). In this study, we take a design-based research approach to redesigning and assessing a simulation-based module, Darwinian Snails, designed to teach evolution by natural selection.

This research builds on a prior study that Abraham et al. (2009) conducted on the original Darwinian Snails module (Herron et al. 2014). Abraham and colleagues found that the module was effective at teaching how natural selection works and overcoming some student misconceptions (Abraham et al. 2009), but it also uncovered some weaknesses in the design of the module. In that study, the authors identified two specific areas in which the Darwinian Snails module could be improved. First, they suggested that it could better increase student understanding by explicitly contrasting the key concepts of natural selection to common misconceptions about the topic, a strategy that had proven effective in other studies (e.g., Jensen and Finley 1996; Robbins and Roy 2007; Kalinowski et al. 2013). Second, they recognized weaknesses in the section of the module focused on mutation as the source of trait variation and suggested that it may inadvertently reinforce the misconception that mutation is induced by the presence of a selective agent in the population, and that mutation is directional (i.e., biased towards "advantageous" mutations). They suggested revising the section on mutation to more directly confront the misconception that mutations are always adaptive.

In this paper, we discuss our approach to redesigning the Darwinian Snails module. Specifically, we focus on the revisions designed to address the weaknesses identified in the previous study (Abraham et al. 2009), and we assess the effectiveness of the redesigned module. The original version of the module used in the study by Abraham et al. (2009) (here referred to as the workbook version) involved onscreen simulations with an accompanying paper workbook that contained instructions and open-response questions. We revised the module by first identifying the key concepts we wanted to teach and the

misconceptions we wanted students to replace or transform into scientifically supported conceptions (Kalinowski et al. 2013). We then transformed the module into a "tutorial" format, with all instructions onscreen, along with multiple-choice and other forced-response assessments that provide immediate feedback, enabling the material to more specifically address common misconceptions and reinforce key concepts (we refer to this as the tutorial version). In addition to describing the redesign, we present two studies assessing the tutorial module. In study 1, we examine the effectiveness of the tutorial using a pre-post instrument aligned to the key concepts of natural selection that we identified as learning outcomes of the module. In study 2, we compare performance on a summative assessment of students who completed either the workbook or tutorial versions of the module.

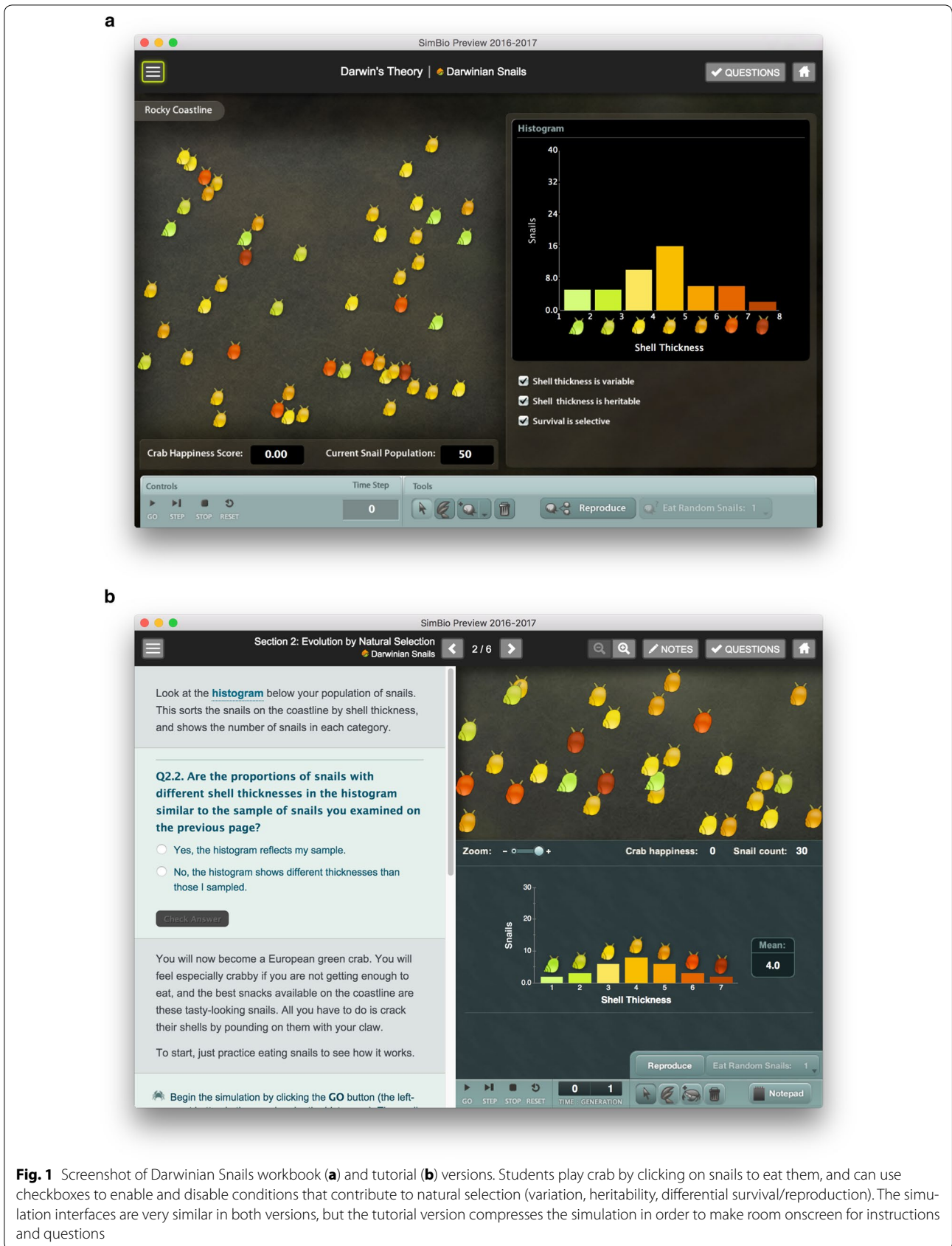
Approach

Design-based research is a methodological approach for the study of learning through the systematic design and analysis of instructional strategies and tools (DBR Collective 2003). Easterday et al. (2014) describe six phases of DBR, in which designers focus the problem, understand the problem, define goals, conceive the outline of a solution, build the solution, and test the solution, and then iterate the process (p. 319). The process involves developing a theoretically driven design that is based on understanding of the content, in this case, concepts and misconceptions around natural selection. Each iteration of the Darwinian Snails module was based on research on how students learn (and don't learn) concepts of natural selection. We implemented each design to examine how it worked and then evaluated and reflected on the process. This was an iterative cycle of design-test-redesign. This paper presents two rounds of design-based research (study 1 and study 2).

Description of module

The Darwinian Snails module

Both the workbook and tutorial versions of the Darwinian Snails module share a common instructional design; we will refer to this as the generic Darwinian Snails module. The module includes a series of interactive simulations that allow students to make predictions on the effect of changing conditions, test these predictions, make observations, and draw inferences about the conditions that lead to evolution by natural selection based on their tests (see Fig. 1 for example screenshots from each version). The module uses Robin Seeley's research on the effect of European green crab predation on the evolution of shell thickness in periwinkle snails (Seeley 1986) to illustrate the process of natural selection. The module has



a number of sections (Table 1), of which the middle ones (1–5 in the workbook; 2–4 in the tutorial) are the most relevant to this study (see Abraham et al. 2009, for a full description). These middle sections begin with students playing the role of a crab preying on snails to see evolution by natural selection in action (Fig. 1). They continue by having students sequentially “turn off” variation, heritability, and differential survival based on shell thickness, in order to investigate how each factor is required for evolution by natural selection to occur. They finish with students exploring the role of mutation as a source of variation in snail shell thickness by disabling and enabling mutations in the presence of the crab predator to see whether mutations are affected by the environment and are biased in the direction that would be selectively advantageous.

The workbook version of the Darwinian Snails module

In the original workbook version, students are provided with a paper workbook that directs them through the different exercises and asks them to observe and interpret the results of the simulations. The workbook contains open response (short answer and essay) questions where

students record their observations and draw conclusions. As mentioned above, Abraham et al. (2009) investigated the effectiveness of this workbook version of Darwinian Snails as a tool for teaching about evolution by natural selection. They found that while effective for helping students overcome some misconceptions, the tutorial’s treatment of mutation was insufficient, and that the module could more explicitly contrast key concepts with misconceptions.

The redesigned tutorial version of the Darwinian Snails module

In 2013, we created a new “tutorial” version by converting the workbook version of Darwinian Snails to an onscreen tutorial format. Since the introduction of the tutorial version, most instructors who adopt the module for use in their courses choose to use the tutorial version, but the workbook version continues to be used in some courses.

The tutorial version has three major differences from the workbook version of the module, two of which were designed to directly address the recommendations of Abraham et al. (2009). We describe these changes below.

Table 1 Outline of the sections of the Darwinian Snails modules in both versions

Exercise name in workbook version	Section name in tutorial version	Description (tutorial version)
Prelude	1. Snail shells	Introduces students to the study system and to histograms as a way of graphing data to visualize the distribution of trait variation in a population
1. A model of evolution by natural selection	2. Evolution by natural selection	Students play the role of a crab preying on snails. They discover that thinner-shelled snails are easier to eat and go through several rounds of selection and reproduction to see a shift in the trait distribution in the population, thus visualizing natural selection in action
2. The requirements for evolution by natural selection 3. Darwin’s theory of evolution by natural selection	3. Requirements for natural selection	Students turn off three conditions for evolution by natural selection in turn to see the effect of each: variability; heritability; and differential survival. In turning off differential survival, students also see genetic drift in the small population. Mutations are disabled in this section
4. The source of variation among individuals 5. What makes populations evolve?	4. The source of variation	The students can now see the effect of mutations in the population. Starting with a snail population with reduced trait variation, students “turn on” mutations and examine parent/offspring combinations and histograms to see whether mutations are directional in the presences and absence of crab predation
6. Challenge: evolution by natural selection in flat periwinkles	5. Testing for natural selection 6. Extension: experimenting with snails	Data from Seeley (1986) are shown and students are asked to interpret the data Students are asked to devise their own experiments to determine whether the conditions for evolution by natural selection are met in a snail population

Rows correspond to sections in the tutorial; descriptions are of tutorial sections. The overall flow of the module is the same in both versions, but the way the material is divided into sections differs, so that some sections in the tutorial include material from more than one section of the workbook, and vice versa. In the final dataset discussed in this study (study 2), the tutorial version no longer included the Extension Sect. (6)—that material was moved to a separate module that was not part of this study

Structural changes

The first major change was the conversion of the module to tutorial format. This consisted of moving the instructions for the task sequence and manipulations of the simulations from the printed workbook to onscreen, alongside the simulation (Fig. 1). In addition, in place of the 31 open-response questions in the printed workbook, the tutorial version includes 31 forced-response questions (the forced-response questions were originally derived from the open response questions, but we also developed new questions; thus, there is not a one-to-one correspondence). Seven of the forced-response questions ask students to make predictions about the outcome of the simulation, and no feedback is provided once they choose a prediction. The rest (24) of the questions, however, provide immediate feedback and thus serve as formative assessment by providing students with feedback on their understanding. Students can keep answering the questions until they get the answer correct.

Contrast of key concepts and related misconceptions

The second major change was to directly contrast key concepts and related misconceptions in the module (as suggested by Abraham et al. 2009). In making the revision, we refined the learning outcomes by identifying six key concepts and six

misconceptions targeted in Darwinian Snails (Table 2). These concepts and misconceptions had been identified in the previous study (Abraham et al. 2009) and in other studies of student thinking about natural selection (Gregory 2009; Bishop and Anderson 1990; Nehm and Reilly 2007), although they are often identified by different labels.

We include several task sequences where we first ask students to make a prediction about the outcome of a particular instance of the simulation (after changing one of the initial conditions), then instruct them to run the simulation, observe the outcome, and then to respond to a reflection question. The reflection questions ask students to reflect on their original prediction and why their prediction may or may not have been supported by the simulation results. Prediction questions did not provide feedback, but the reflection questions do.

For instance, before removing trait variation from the simulation, students are asked to predict what will happen (bold text indicates the correct response):

Question: Do you think this population of snails will evolve as predators start eating them? Why or why not?

- A. Yes, the population will evolve toward thicker shells, because the snails need protection against predatory crabs.
- B. Yes, the population will evolve toward thicker shells, because some snails will grow a thicker shell in order to survive.
- C. Yes, the population will evolve toward thicker shells, because snails in each new generation will have slightly thicker shells than the last one.
- D. **No, the population will not evolve toward thicker shells, because there is no variation in shell thickness.**

As a prediction question, students did not receive feedback on the correctness of their response. After removing trait variation, running the simulation, and observing that evolution by natural selection cannot occur without trait variation, students are then asked the following reflection question:

Question: Many students predict that the snail shell thickness would still evolve even without variation because the snails need protection against predatory crabs. Why didn't you see this in your experiment?

- A. **Without variation in shell thickness, the snails that survive are no different than the ones that are eaten, and so the next generation's shells will always be the same thickness as the previous generation's.**
- B. The snails in my experiment were able to survive even without thicker shells, so they didn't need to

Table 2 Six key concepts and six target misconceptions

Key concept	Description
Evolution by natural selection	Change in relative frequency of trait in a population
<i>Need to change</i>	Populations/organisms change traits because they need to
<i>Individuals evolve</i>	Evolution occurs because individuals change their traits
<i>Gradual population change</i>	Evolution occurs via gradual change in whole population (i.e., all individuals from one generation to the next)
Variation	Trait variation in a population is necessary for natural selection
Role of mutation	Mutation is a source of trait variation
Mutations random	Mutations are random and unrelated to selective pressure
<i>Adaptive mutation</i>	Mutations are adaptive responses to the environment and are biased towards advantageous mutations
Heritability	Heritability of a trait is necessary for natural selection
<i>Beneficial traits</i>	Offspring inherit only beneficial traits
<i>Acquired traits</i>	Acquired characteristics are inherited
Differential fitness	Differential survival and/or reproduction necessary for natural selection

Key concepts in bold, and misconceptions following their most closely associated key concept in italics

evolve. If they had needed thicker shells, I would have seen evolution in shell thickness.

- C. Without variation in shell thickness, evolution toward thicker shells will take longer; the experiment only lasted one generation, so there wasn't enough time for evolution to take place.

The feedback for reflection questions was designed not only for students to learn whether their answer is correct, but also to encourage them to think about why their answer was correct or incorrect. In the revisions, we created distractors aligned with our target misconceptions and structured feedback to those responses to explicitly confront those misconceptions. Students are prompted to try again if they initially choose the wrong answer. For instance, the feedback for choice B in the question above reads:

That's not correct. Snails would certainly benefit by having thicker shells, but just because they would benefit does NOT mean that the process of evolution can provide thicker shells! Try again.

Redesign of section on the role of mutation

The third major change was to restructure the student task sequence in the section of the module in which students explore the role of mutations (see Table 1, Sect. 4). In the new sequence, we attempt to confront the misconception that mutations are directional/adaptive, providing students with the opportunity to directly observe that mutations are random and that this randomness is not affected by the presence or absence of a selective agent. In our first phase of revisions for the tutorial version (tested in study 1), we expanded the mutation section of the module to include five multiple-choice questions with immediate and specific feedback (formerly there were three open-response with no feedback). We designed the multiple-choice questions to contrast the key concept of random mutation with the misconception of adaptive mutation. We also changed the variation in the starting population of simulated snails. In the workbook version, the starting population in this section displayed a reduced range of the key trait (shell thickness) compared to previous sections. In the revised tutorial version, the starting population displayed just a single value of the key trait, making it more obvious that mutations occurred in both adaptive and maladaptive directions.

Testing during the first revision

As part of the design-based research process, we tested the revised tutorial version with 31 students. The students were recruited from introductory biology classes at public and private higher education institutions in the northeast United States. We observed students as

they progressed through the module, stopping them at many specific points to probe with questions to uncover the thinking behind their actions and their reasons for choosing specific answers. We continually modified the module based on our observations and student feedback during the revision process and tested until we were satisfied that we had addressed common points of confusion about the module content or the interface.

The second phase of revisions

In a second phase of the design-based research process (tested in study 2), we replaced one of the multiple-choice questions in the mutation section with two less-constrained questions (Fig. 2). This question format, which we call LabLibs, is modeled after fill-in-the-blank questions. Students are given a sentence with blanks, and must fill each blank by choosing options from a set of choices in a drop-down menu. Because of the increased number of possible answers, this format pushes students towards constructing an answer and discourages use of test-taking strategies such as process of elimination, compared to a multiple-choice format. The first of the LabLibs questions asked students to predict whether mutations will be directional in the presence of crabs (providing feedback only on the consistency of their answer), and the second LabLibs question asked students to reflect on the results they observed, including mutations and the presence of crabs, after repeatedly running the simulation.

In addition to the three major revisions above, two of which directly addressed the weaknesses identified by Abraham et al. (2009), the tutorial version included a large number of smaller changes, which may have contributed to findings discussed in this paper. These are listed in Additional file 1: Table S1, with short descriptions of each change. There were additional changes made after classes used the tutorial version for one semester (after study 1 but prior to study 2). These are listed in Additional file 1: Table S2.

Study 1

In study 1 we examined the first phase of redesign of the module (converting it to the tutorial version, along with the associated changes described above). The research questions that guided this study were:

- (1) Do we see an improvement in understanding of the process of natural selection after using the tutorial version of the module, as measured by the pre and post assessment?
 - a. Is there a difference in understanding between beginner and advanced students?

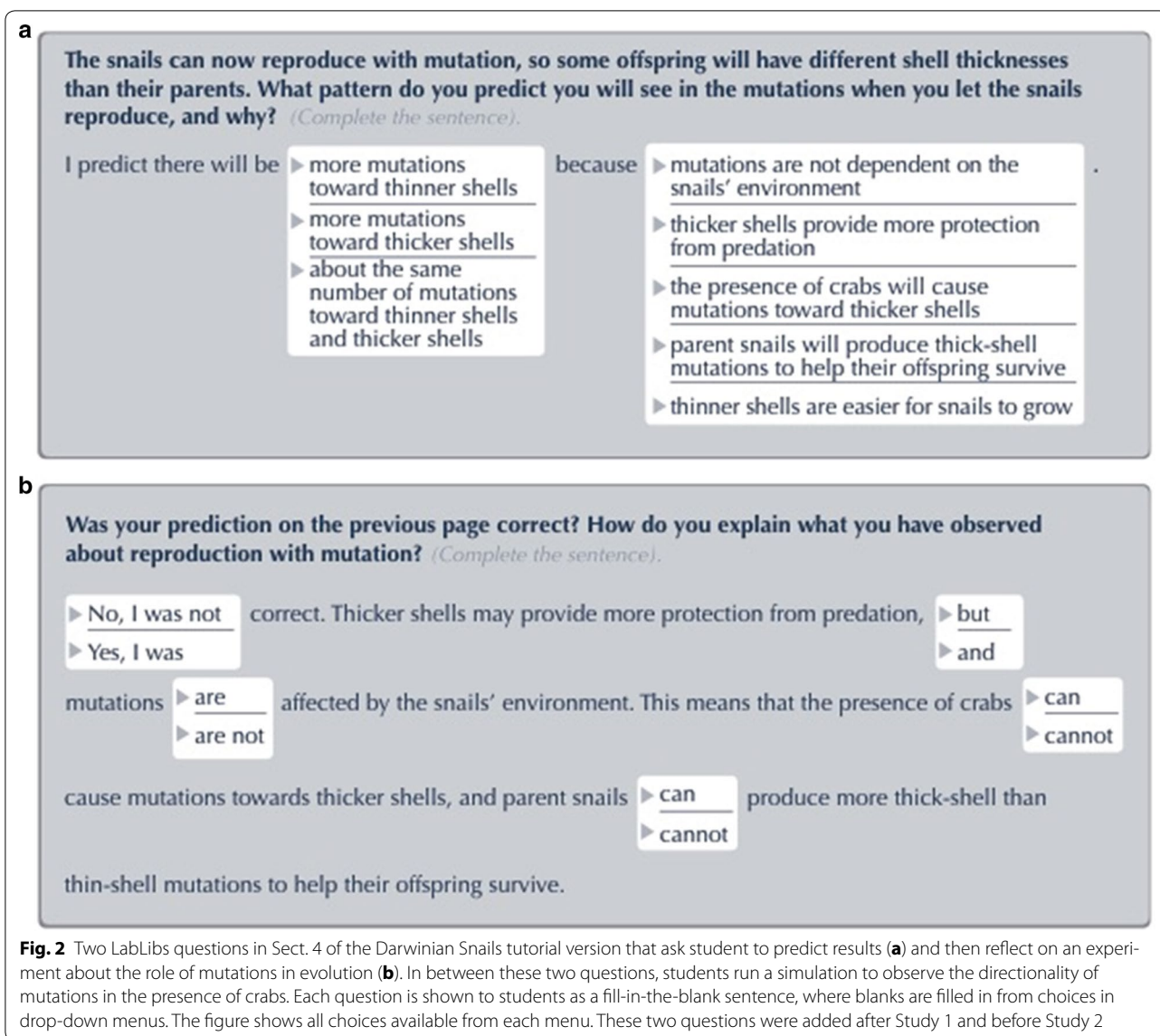


Fig. 2 Two LabLibs questions in Sect. 4 of the Darwinian Snails tutorial version that ask student to predict results (a) and then reflect on an experiment about the role of mutations in evolution (b). In between these two questions, students run a simulation to observe the directionality of mutations in the presence of crabs. Each question is shown to students as a fill-in-the-blank sentence, where blanks are filled in from choices in drop-down menus. The figure shows all choices available from each menu. These two questions were added after Study 1 and before Study 2

- (2) Do we see an increase in key concept understanding and decrease in misconception use after using the tutorial version of the module?
- (3) How do these findings compare to the previous findings on key concepts and misconception use in the workbook version?

Study 1 methods

Sample

We recruited college biology courses to participate in the study through a combination of asking professors already using the Darwinian Snails module in their class and recruiting classes through a series of webinars that were advertised to SimBio's mailing list of biology faculty. Professors who agreed to have their course participate in the

research were offered either a 20% discount on the cost of the module or, in the case of professors new to using SimBio modules, free use of the module. Costs for SimBio modules were generally paid for directly by students as part of their required materials for the course, but in some cases were paid from a department budget.

Our sample included 1362 students in 13 courses (Additional file 1: Table S3). Students were 18 years of age or older and enrolled in introductory and upper-level biology courses. As in previous research (Abraham et al. 2009), we refer to students in the introductory courses (100 level) as "beginner" and those in courses requiring prior coverage of ecology and evolution (200 or 300 level) as "advanced." For students who reported their gender,

767 identified as female, 436 identified as male, and 5 identified as transgender.

The Committee on the Use of Humans as Experimental Subjects, the institutional review board at the Massachusetts Institute of Technology in Cambridge, MA, approved this study before data collection (COUHES #1206005102), and for each of the classes whose data we used, we also received approval from the IRBs of their institutions (they either chose to review and approve the study or accepted the approval of MIT COUHES).

Description of assessment instruments

To measure learning gains, we administered a pre/post-test embedded in the module. We developed the assessment, selecting items from three different published assessments (Bishop and Anderson 1986; Anderson et al. 2002; American Association for the Advancement of Science 2013) that aligned with our key concepts. We also classified each answer option as expressing one of our key concepts or target misconceptions (Additional file 1: Table S4). In some cases, a given misconception was expressed in more than one answer option for the same item. Some distractor answer options could not be classified with a target misconception—we classified those as “misunderstanding of [KC]” (KC being whichever key concept was assessed by the question). These distractors were generally selected less frequently than other answer options on the same item. Our instrument includes 14 multiple-choice items. Internal consistency was measured using Cronbach’s α (DeVellis 2003). Cronbach’s α was 0.75 for both the pre and post-test.

Multiple-choice questions are limited in their ability to elucidate true student understanding of difficult concepts (Resnick and Resnick 1992; Nehm and Schonfeld 2008; Quellmalz and Pellegrino 2009). Open-response items, such as short answer or essay questions, are often more revealing (Nehm and Schonfeld 2008). In the pre and post-test, we also included an open-response question from the ACORNS instrument (Nehm et al. 2012). ACORNS includes a variety of items that ask the same question but vary the organism and trait involved. We used the ACORNS item: “Orchids are a type of flowering plant. How would biologists explain how a living orchid species lacking leaves evolved from an ancestral orchid species that had leaves?” This item contains two elements (a plant example and loss of function) that Nehm and colleagues (Nehm et al. 2012) suggest are likely to be challenging for students.

Using the original Abraham et al. (2009) assessment instrument in this study would have facilitated direct comparison with the results of that study, but we decided against using that assessment. Our redesign of the tutorial version involved refining the learning goals (Table 2),

which did not clearly align with the items on the previous assessment. Thus, as described above, we assembled a new assessment that measured our key concepts and included our targeted misconceptions in answer choices.

Analysis

Analysis of performance on multiple-choice instrument

We calculated item difficulty on each test by dividing the number of correct responses for each item by the total number of responses for that item (Crocker and Algina 1986).

Performance scores were gathered at repeated time points longitudinally on students who were also nested within classes; therefore, a hierarchical or multilevel model (HLM) was required to account for lack of independence between scores. The intraclass correlation (ICC) indicates the proportion of the variance explained by the grouping structure of the population. We used a three-level model where level one accounted for the repeated measures of performance, level two accounted for student-specific correlation between these scores ($n=1322$, $ICC=0.19$), and level three accounted for additional class effects ($n=13$, $ICC=0.34$). Therefore, *time* is a level-one factor and *type of class* is a level-three factor that indicated whether the course was at an advanced or beginner level. To capture change in student understanding of natural selection, we were most interested in the effect of *time* (change from pre to post-test). As in prior research, we were also interested in whether the *type of class* had an effect overall (five advanced vs eight beginning courses) or if the *type of class* influenced the change (interaction between time and type of class). All analyses were performed in R 3.3.3 (R Core Team 2017) utilizing the ‘lme4’ package (Bates et al. 2015).

We used Cohen’s *d*, a standardized measure of the differences between the means, to calculate the effect size. Effect size provides a description of magnitude of the observed effect that is independent of sample size (Fritz et al. 2012). We calculated normalized gain scores between the pre-test and post-test, which measures the amount of improvement demonstrated by students relative to the amount that they could have improved on the assessment instrument (Hake 1998; Theobald and Freeman 2014).

Analysis of misconceptions using multiple-choice responses

In addition to comparing students’ total correct scores on the pre and post-test, we assessed students’ misconceptions on the pre and post-test multiple-choice instrument by examining their selection of the distractors (incorrect responses) that we had classified as expressing one of our target misconceptions. In a forced-choice instrument like this one, an increase in key concepts necessarily means a

decrease in misconceptions; however, because each item on the assessment may have incorrect responses that represent different misconceptions, this analysis complements the key concept analysis by examining which misconceptions changed. For example, for item 13 on the assessment, we classified the correct answer as corresponding to the key concept Evolution by Natural Selection, and the three distractors were classified as aligning with three different misconceptions: Individuals change, Adaptive mutation, and Acquired traits (see Additional file 1: Table S4). Because most of our target misconceptions appeared in more than one item (between 1 and 5 items), for each student, we totaled the number of distractors they selected that included a given misconception and standardized the counts by dividing by the total possible selections (e.g., for a misconception that appeared as a distractor in 3 different items we divided by 3). Thus, the possible standardized count for each misconception ranges from 0 (for students who never chose a distractor with that misconception) to 1 (for students who chose all instances of distractors with that misconception).

Because the standardized misconception counts were not normally distributed, we used a nonparametric Wilcoxon signed-rank test to compare counts between the pre and post-tests. We adjusted alpha for multiple comparisons using the Bonferroni method, adjusting the critical α value to 0.01.

Analysis of performance on open-response instrument

Eight hundred fifty-one students completed the ACORNS open-response item on both the pre and post-test. We used the online grader EvoGrader (<http://www.evograder.org/>, Moharreri et al. 2014) to score their responses for the presence of concepts and misconceptions. There are some differences between the key concepts and target misconceptions we focused on and the concepts and misconceptions (naïve ideas) that EvoGrader assesses. EvoGrader scores for two of our key concepts (heritability and differential fitness), one concept which is a combination of two of our key concepts (variation and role of mutation), as well as two concepts that we do not target in the module (competition and limited resources), and one that we give only limited coverage to (genetic drift). It scores for two of our misconceptions (Need to change and Individuals evolve), as well as one we do not include (use/disuse). We limited our analyses to the three EvoGrader key concepts and two misconceptions that we had targeted in our redesign of the Darwinian Snails module that EvoGrader scores (see example of a student response and EvoGrader score in Additional file 1: Table S5).

We analyzed the ACORNS scores in two ways. First, we compared the average total number of key concepts (out of 3) and misconceptions (out of 2) between the pre and post-test. These data were not normally distributed, so we used nonparametric Wilcoxon signed rank tests to compare totals between pre and post. We calculated the effect size by calculating r using the Z value, where we divided Z by the square root of N (Rosenthal 1994; Fritz et al. 2012). Second, we used a McNemar's test for each key concept or misconception to compare the number of students who did not include it on their pre-test but did on their post-test (which translates into improved performance for key concepts, but decreased performance for misconceptions) to the number of students who included it on their pre-test but did not on their post-test (a decrease for key concepts, improvement for misconceptions). The McNemar's test is essentially a χ^2 test for paired data (McNemar 1947). We used the Bonferroni method to correct for multiple comparisons, adjusting the critical α value to 0.01.

Study 1 results

Student understanding in the tutorial module increased as evidenced by responses to multiple-choice questions

On average, student performance on the multiple-choice questions significantly improved after working through the module (Table 3). Student performance increased from a mean pre-test score of 66.17 (SD = 19.45) to 78.28 on the post-test (SD = 16.92), for a mean gain of 12.11 (SE = 0.41). Over half of the variance among scores is attributable to the hierarchical nature of the data, in other words, the fact that students were nested in classes (variance estimates: student level = 170.36, class level = 73.02, residual = 113.40). The normalized gain score was 0.36 and the effect size was 0.64, a medium effect size for education interventions such as this one (Cohen 1988). The full model taxonomy can be found in the Additional file 1: Table S6.

Advanced students learn as much as beginner students from the tutorial module

As in previous research (Abraham et al. 2009), we compared results between students in advanced level classes and students in beginner level classes (*type of class*). Advanced students did not perform any differently than beginner students: Both beginning and advanced students had similar pre-test scores and showed similarly sized improvements between pre and post-test (see Additional file 1: Table S7 and Table 3 below). There was not an interaction between *type of class* and *time* ($X^2(1) = 0.464$, $p = 0.496$) or a main effect of *type of class* ($X^2(1) = 0.621$, $p = 0.431$).

Table 3 Study 1—Summary of hierarchical linear model results of pre and post-test data

Q: Do students show improvement on assessment after using the tutorial version of the module?

A: Yes. Students had a mean gain of 12.11 points after using the tutorial version of the module

Score ~ time		
Fixed predictor	Estimate ± SE	p value
Intercept	64.01 ± 2.55	< 0.0001
Time	12.11 ± 0.41	< 0.0001

Q: Do advanced students perform better than beginner students on the assessment?

A: No. Advanced students did not perform better than beginner students

Score ~ time + type + time * type		
Fixed predictor	Estimate ± SE	p-value
Intercept	62.25 ± 3.42	< 0.0001
Time	12.30 ± 0.50	< 0.0001
Type of class	4.25 ± 5.33	0.445
Time * type of class	- 0.61 ± 0.89	0.496

Both models include random factors for student (level 2) and class (level 3) to account for the lack of independence between scores. Full model taxonomy included in Additional file 1: Table S6; means for beginning and advanced classes shown in Additional file 1: Table S7 (2644 observations on 1322 students within 13 classes)

Student misconceptions in the tutorial module decreased as evidenced by responses to multiple-choice questions

Comparing the standardized misconception counts from the multiple-choice questions between pre and post-tests, there were significant decreases in all 6 of our target misconceptions (Fig. 3; Wilcoxon signed-rank test; $p < 0.0001$ for all comparisons). In particular, one large change from pre to post-test was in the Adaptive mutation misconception, which was the one misconception that did not see a significant improvement in the previous study (see Fig. 6 in Abraham et al. 2009). These reductions in misconceptions between pre and post-tests could have occurred because students shifted their responses from one misconception on the pre-test to a different misconception on the post-test. To see whether this was true, we calculated the percentage of students who chose different responses on the pre vs post-test on the three questions that included the Adaptive mutation misconception, and if they changed to the correct answer or to a different incorrect answer (Additional file 1: Table S8). Of those who responded differently on the post-test, most shifted from incorrect to correct responses, rather than choosing answers associated with other misconceptions.

Student understanding increased and misconceptions decreased as evidenced by response to an open response question (ACORNS)

Changes in student understanding as measured by the ACORNS open-response question broadly agreed with the results from the multiple-choice instrument. When looking at the key concepts and misconceptions captured by EvoGrader that corresponded to those targeted

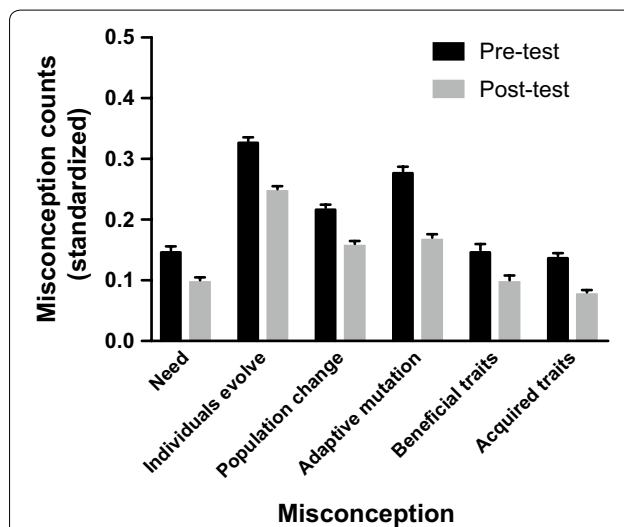


Fig. 3 Standardized misconception counts calculated from multiple-choice items for students using the tutorial version of Darwinian Snails on pre and post-tests in study 1. Misconception counts significantly decreased from pre to post-test for all 6 target misconceptions

in the module, we saw improvement in student expression of both key concepts and misconceptions. Students used significantly more key concepts in the post-test (Mean = 1.49 ± 1.06 , $Mdn = 1$) than in the pre-test (Mean = 0.89 ± 0.89 , $Mdn = 1$); $Z = 14.21$, $p < 0.001$; with an effect size (r) of 0.49. On the pre-test, ~39% of students had zero key concepts in their responses, and only ~22% used two or three concepts; in contrast, on the post-test, ~22% had 0 key concepts, and ~49% used two or three concepts.

Students used significantly fewer misconceptions on the post-test (Mean = 0.25 ± 0.48; *Mdn* = 0) than in the pre-test (Mean = 0.35 ± 0.55; *Mdn* = 0); $Z = 4.60$, $p < 0.0001$; effect size (r) of 0.16. Most students used zero misconceptions in their answers on both pre- and post-tests, but the percent using one or two misconceptions did decrease from pre- to post-test (from ~31 to ~23%).

While about 40% of students showed no change in the total number of key concepts they used in their answers, over 48% increased the number of key concepts they used in responding to the ACORNS question from pre-test to post-test (Fig. 4); about 12% used fewer key concepts on the post-test than on the pre-test.

McNemar’s tests demonstrate that for each of the three key concepts, there was a significant difference in the number of students who improved on their expression the concept (i.e., they did not include it in their pre-test but did in their post-test) compared to the number who declined (i.e., included it in their pre-test but not in their

post-test; Table 4). For each of the two target misconceptions, there were similar significant improvements (that is, more students used it in their pre-test but not in their post-test than vice versa).

For the key concepts and misconceptions that were scored by EvoGrader but were not targeted in our revisions, there was only one significant difference (an improvement), in the expression of the Limited resources key concept (Additional file 1: Table S9). Other than the Limited resources concept, these key concepts or misconceptions were expressed at very low frequencies on both pre and post-tests.

How do these findings compare to the previous findings on key concepts and misconception use in the original version?

Table 5 presents a comparison between the original workbook version of the lab (Abraham et al. 2009) and the revised tutorial version. We found evidence for a decrease in the use of the Adaptive mutation misconception for students using the revised tutorial version. Using a different assessment, Abraham et al. (2009) did not find evidence for a decrease in use of this misconception for students using the original workbook version.

Study 2

In study 2, we directly compared student learning between the workbook version and the revised tutorial version of the Darwinian Snails module, based on their performance on a summative assessment that is included at the end of both versions of the module. In particular, we were interested in evidence that the revisions to the tutorial version section on mutation reduced students’ use of the adaptive mutation misconception. The research questions that guided this study were:

- (1) Do students who complete the revised tutorial version perform differently on a summative assessment of concepts covered in the module than their peers who used the original workbook version?

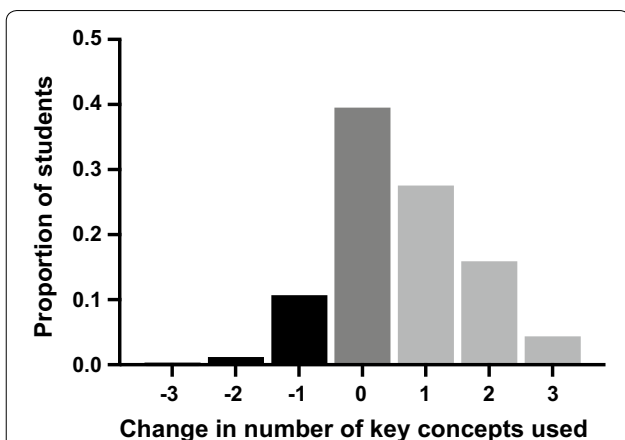


Fig. 4 Change in # of key concepts used in response to ACORNS question from pre to post-test in study 1. 40% showed no change (dark gray bars); 48% showed an increase in key concepts used (light gray bars). This only includes the 3 key concepts scored by EvoGrader that are part of our target key concepts (Variation, Heritability, and Differential Survival)

Table 4 Changes in key concepts and misconceptions as measured by the ACORNS item

	Proportion expressing on pre	Proportion expressing in post	McNemar’s χ^2	p value* (df = 1)
<i>Key concept</i>				
Variation	0.34	0.53	82.4	<0.0001
Heritability	0.13	0.35	127.2	<0.0001
Differential fitness	0.42	0.62	97.7	<0.0001
<i>Misconception</i>				
Need to change	0.25	0.20	9.4	<0.01
Individuals evolve	0.10	0.05	16.4	<0.0001

*Correcting for multiple comparisons, critical value is 0.01. At this level, all comparisons are significant

Table 5 A comparison of student improvement on misconceptions

Misconception	Significant improvement in workbook version?	Significant improvement in tutorial version?
Need to change	Yes	Yes
Individuals evolve	Yes	Yes
Gradual population change	Yes	Yes
Adaptive mutation	No	Yes
Beneficial traits	Not assessed	Yes
Acquired traits	Not assessed	Yes

Comparison of student improvement on misconceptions when using the workbook version of the Darwinian Snails module (as measured by Abraham et al. 2009) or using the tutorial version (data from study 1; Fig. 3; Table 4). Because the assessment instruments were different, only a qualitative comparison is shown

- a. Is there a difference in performance between beginner and advanced students?
- (2) Is there a difference in the prevalence of the adaptive mutation misconception between students using the tutorial version and students using the workbook version?

Study 2 methods

Sample

Our sample consisted of 2605 students spread across 38 classes (Additional file 1: Table S10). Of these, 1885 students in 18 classes used the tutorial version of the module and 720 students in 20 classes used the workbook version. The tutorial classes were larger on average (average = 105 students) than the workbook classes (average = 36 students).

These data were collected as part of normal classroom use of the Darwinian Snails module tutorial and workbook versions. The assessment we used is integrated into the module and instructors often use scores on the assessment for completion or performance credit for students. After the completion of the semester, we obtained approval from the New England Independent Review Board (NEIRB #120160152) to use the de-identified answer data for research purposes. Because the data were not originally collected for research purposes, we did not collect any demographic information from the students, nor did we recruit classes or offer compensation of any kind.

Description of assessment instrument

After completing the Darwinian Snails module, students using both versions submitted responses to the same ten summative assessment questions. The ten items included four questions taken from the instrument used in study 1 without modification, two questions based on items from that instrument but with some modifications, and four newly written items designed specifically to test material from the module, to help instructors confirm that students had completed the module (Additional file 1:

Table S11). This was a study of opportunity rather than a planned study, and this assessment was not originally intended for research. Thus, neither the newly written questions nor the assessment as a whole were subject to validation beyond internal review within the project team. Internal consistency (Cronbach’s α) was 0.66 for the tutorial version and 0.71 for the workbook version.

Analysis of performance on multiple-choice instrument

Similar to study 1, performance scores were gathered on students who were nested within classes, therefore a hierarchical model (HLM) was required. We used a two-level model where level one accounted for student-specific correlation between these scores, and level two accounted for additional class effects (ICC = 0.17). Therefore, *treatment* is a level-one factor and *type of class* is a level-two factor. To capture whether the revisions to our module increased student understanding of natural selection, we were most interested in the effect of *treatment* (whether they used the tutorial version or the workbook version) on their performance measure (*score*). As in prior research, we were also interested in whether the *type of class* had an effect overall (15 advanced vs 28 beginner courses) or if the *type of class* influenced the *treatment* (interaction between treatment and type of class). All analyses were performed in R 3.3.3 (R Core Team 2017) utilizing the ‘lme4’ package (Bates et al. 2015).

To address research question 2 for study 2, we assessed students’ expression of misconceptions in the classes that used either module version as indicated by their selection of distractor answer options on the graded questions (as in study 1). For this comparison, we focused on the adaptive mutation misconception, because we had specifically targeted this misconception as part of our revisions for the tutorial version. This misconception appeared as a distractor in three different items (one newly written item, one item revised from the study 1 instrument, and one item identical to the study 1 instrument; see Additional file 1: Table S11). We standardized the counts for

each student by dividing by 3; thus, possible standardized misconception counts ranged from 0 to 1. Because the standardized misconception counts were not normally distributed, we used the nonparametric Wilcoxon/Mann–Whitney test to compare counts between students using the workbook vs tutorial versions. We calculated the effect size r by dividing Z by the square root of N (Rosenthal 1994; Fritz et al. 2012).

Study 2 results

There was no statistically significant difference in performance on the summative assessment between students who used the tutorial and those who used the workbook version of the module.

There was not a main effect of *treatment*, workbook vs tutorial version, ($\chi^2(1)=0.03$, $p=0.8738$ (Table 6). In other words, when we accounted for the clustering of students in classes, there was not a significant difference in performance for students who used the tutorial version (Mean=73.12±20.61) and those who used the workbook version (Mean=68.65±22.85) on the summative assessment. About one-fifth of the variance among scores is attributable to the hierarchical nature of the data, in

other words, the fact that students were nested in classes (variance estimates: class level = 81.90, residual = 408.19).

The module version did not have different effects on the performance of advanced vs beginner students

We did not find an interaction between *treatment* and *type of class*, $\chi^2(1)=0.1433$, $p=0.705$ (Table 6), suggesting that there was no interaction between the version of the module used (workbook or tutorial) and the academic level of the students. Students in advanced classes did perform better than students in beginner classes on the summative assessment, averaging 6.44 points higher (SE=3.13), and this main effect *type of class* (Advanced vs Beginner) was significant, $\chi^2(1)=4.20$, $p<0.0001$. However, the difference between academic levels was not affected by module version. The full model taxonomy can be found in Additional file 1: Table S12.

Students using the tutorial version had lower misconception counts for the Adaptive mutation misconception than those using the workbook version

The Adaptive mutation misconception appeared in distractors in three different questions on the assessment. Students in courses that used the workbook version

Table 6 Study 2—summary of hierarchical linear model results of outcomes on the summative assessment

Q: Does treatment (workbook version vs tutorial version) predict student summative assessment score?		
A: No. Students in the two treatments, workbook version and tutorial version, did not differ in their summative assessment performance		
Score ~ treatment		
Fixed predictor	Estimate ± SE	p value
Intercept	69.55 ± 2.32	< 0.0001
Treatment (workbook)	− 0.49 ± 3.24	0.881
Q: Does treatment (workbook version vs tutorial version) predict student summative assessment score controlling for their type of class (Advanced vs Beginner)?		
A: No. There is not an interaction between treatment and type of class		
Score ~ treatment + type of class + treatment * type of class		
Fixed predictor	Estimate ± SE	p value
Intercept	67.80 ± 2.76	< 0.0001
Treatment (workbook)	− 2.20 ± 4.04	0.589
Type of class (Advanced)	5.35 ± 4.79	0.272
Treatment * type of class	2.33 ± 6.50	0.722
Q: Do advanced students perform better than beginner students on the summative assessment, regardless of treatment?		
A: Yes. Advanced students perform better on the summative assessment regardless of which module version they used (workbook or tutorial)		
Score ~ Type of class		
Fixed predictor	Estimate ± SE	p value
Intercept	66.79 ± 1.97	< 0.0001
Type of class (advanced)	6.44 ± 3.13	< 0.05

All models include random factors for class (level 2) to account for the nesting of students in classes (2605 students within 38 classes). Full model taxonomy can be found in Additional file 1: Table S12

chose a distractor expressing the Adaptive mutation misconception more frequently than those using the tutorial version; the standardized score for this misconception was significantly higher for the workbook students (Mean = 0.32 ± 0.32; Mdn = 0.33) than the tutorial students (Mean = 0.23 ± 0.28; Mdn = 0; Z = 5.98, p < 0.0001), r = 0.12. Another way to visualize the data is to compare the proportion of workbook and tutorial students who chose an Adaptive mutation distractor 0, 1, 2 or 3 times (Fig. 5); 28% percent of students using the workbook version selected an Adaptive mutation distractor on two or three different questions, compared to 17% of students using the tutorial version. More than 50% of students using the tutorial version and more than 40% using the workbook version did not choose this misconception distractor on any of the 3 questions. Looking specifically at the distribution of responses to one of these three questions, which was written specifically to assess the Mutations random key concept, workbook students seemed to be more attracted to the response suggesting that mutations are biased in the direction that would convey a selective advantage (Additional file 1: Table S13).

Discussion

We applied design-based research to investigate whether changes to a simulation-based module on natural selection improved undergraduate student learning of major concepts. This process involved developing a theoretically driven design based on understanding of the concepts and misconceptions around natural selection. As described previously, Easterday et al. (2014) outlined six phases of DBR. The first two phases are to focus on the problem and understand it. We did this by following up on an earlier study by Abraham et al. (2009), which found evidence that the original module (with onscreen simulations and a paper workbook) significantly increased student understanding of natural selection, but also identified two areas where the module could be improved. Next, we defined our goals and outlined a solution by identifying where in the module we could make changes. Specifically, we identified six concepts and six misconceptions (see Table 2) to focus on, based on the research conducted on the original module (Abraham et al. 2009) and other research on natural selection (e.g., Kalinowski et al. 2013; Gregory 2009; Nehm and Schonfeld 2008; Ferrari and Chi 1998). Then, we redesigned the module (e.g., we updated the format to a tutorial style with onscreen instructions and immediate feedback to forced-response questions), and tested these changes, as described in study 1 and study 2. In this iterative process, each redesign was based on what we had previously found about how students learned and where they struggled with the concepts of natural selection from the

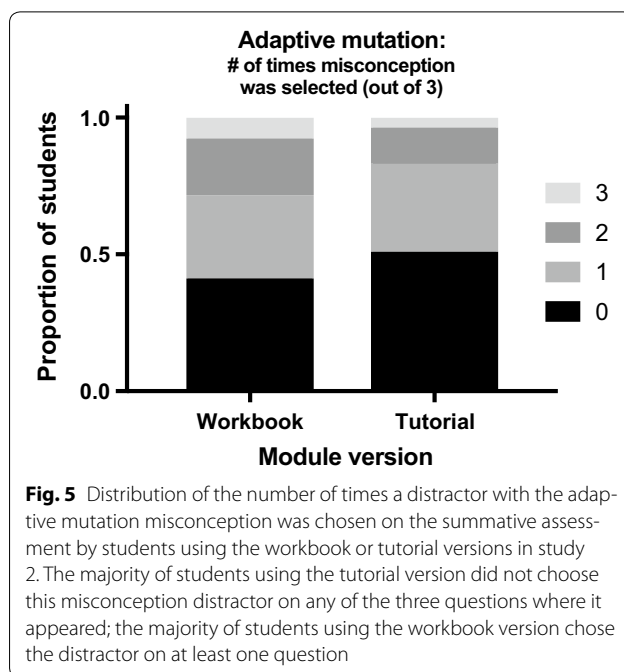


Fig. 5 Distribution of the number of times a distractor with the adaptive mutation misconception was chosen on the summative assessment by students using the workbook or tutorial versions in study 2. The majority of students using the tutorial version did not choose this misconception distractor on any of the three questions where it appeared; the majority of students using the workbook version chose the distractor on at least one question

module. Our findings suggest that the revised tutorial version performs just as well as, and in some ways, even better than, the original workbook version, demonstrating the effectiveness of the iterative approach of design-based research—using knowledge gained from assessing pedagogical tools to improve those tools.

Results from study 1

In our first phase of design-based research, we found that students who used the revised tutorial module showed an increase in understanding of natural selection as evidenced by their gain on pre-post test items, including both multiple-choice questions (Table 3) and an open-response question (Fig. 4). The multiple-choice items were specifically targeted towards the key concepts and used target misconceptions in distractors.

The use of misconceptions as distractors in the multiple-choice questions allow for more complete assessment of student conceptual knowledge (Anderson et al. 2002); based on their answer selections, students who used the tutorial showed a decrease in selection of misconceptions from pre to post-test (Fig. 3). In particular, we found a significant decrease in all six of our target misconceptions. While we cannot make a direct comparison to the earlier study (Abraham et al. 2009), our results do show a significant decrease in the Adaptive mutation misconception, which did not show a significant change in the previous study. We also found significant decreases in two misconceptions about inheritance that were not assessed in that previous study—Beneficial traits and Acquired traits

(Table 5). Thus, while it is possible that the assessment tool or analyses in Abraham et al. (2009) were not sensitive enough to detect a change in this misconception, it appears clear from our results that the revised tutorial version of the module does improve student performance on the common Adaptive mutation misconception.

Because research on natural selection assessments has found that the format of the question plays an important role in students' expression of natural selection concepts and misconceptions (Nehm and Schonfeld 2008), we also used one of the open-response questions from the ACORNS instrument (Nehm et al. 2012). Using this open-response assessment, we found similar results to the multiple-choice assessment: Students' expression of key concepts increased (Fig. 4) and their use of misconceptions in their explanations decreased after completing the module (Table 4). The gain in number of key concepts used in the externally developed open-response assessment in this study (Fig. 4) provides additional support for the efficacy of this revision of the module.

In a study conducted concurrently with study 1, Pope et al. (2017) compared the tutorial version of the module to a physical simulation of natural selection using a split-class design in a large-enrollment introductory biology lab class (where half the sections completed the Darwin Snails tutorial and half completed a physical simulation). Using the same pre-post multiple-choice assessment we used in study 1, Pope et al. found significant gains from pre to post, but no difference in gains between the virtual (Darwinian Snails) simulation and the physical simulation, providing further evidence that the tutorial version is an effective tool to help students learn natural selection.

Results from study 2

In our second phase of design research, we compared the performance of students who used the original workbook module to their peers who used the revised tutorial module, on a short end-of-unit summative assessment. We did not find any significant differences in the outcome between the two treatments (Table 6), suggesting that both modules are equally effective for student learning of key concepts. We also found that students who used the revised tutorial version had lower misconception counts for the Adaptive mutation misconception than those who used the original workbook (Fig. 5), in line with our findings from study 1 suggesting that the revisions to the tutorial were effective in helping students overcome this misconception.

Adaptive mutation misconception

A common misconception about evolution by natural selection is that mutations are adaptive responses to the

environment and are biased towards advantageous mutations. The misconception suggests that selection pressure or environmental conditions, rather than random mutation and genetic recombination through sexual reproduction, are the causes of new trait variation in a population. In an open-response assessment in the previous study by Abraham et al. (2009), this was the most commonly expressed misconception on both the pre and post-test. This is not surprising given that processes involving randomness are difficult for students to understand (Ferrari and Chi 1998; Meir et al. 2007; Garvin-Doxas and Klymowsky 2008; Price et al. 2016).

In addition to being the most common misconception in the previous study, they also found that the adaptive mutation was the only misconception that did not significantly decrease after using the original workbook version of the module. In the section on mutation in the workbook version of the module, the initial population displayed a reduced range of the key trait (shell thickness), which is realistic but makes new thicknesses arising through mutation difficult to detect. Abraham et al. (2009) suggested design changes to this section for future versions. In the revised tutorial version, we confronted the Adaptive mutation misconception head-on by creating an unnatural population of snails with no initial variation, which allows students to directly see the mutations and observe that they occur in both adaptive and maladaptive directions; this approach may have contributed to an improvement in students' understanding of the random nature of mutation. We found evidence of this improvement in both phases of our design-based research (study 1—Table 4; study 2—Table 5). This is similar to findings that show simulations can be useful for teaching evolution because randomness is difficult for students but can be observed more readily in simulated populations than real ones (Soderberg and Price 2003; Smetana and Bell 2012). This example of the re-design demonstrates how using design-based research contributes to understanding of how to help students overcome challenging misconceptions around adaptive mutations.

Advanced vs beginner students

Previous studies have suggested that students need to learn about natural selection multiple times and at multiple academic levels, as it is not something they are likely to master by seeing it once in introductory biology (Nehm and Reilly 2007; Kalinowski et al. 2013; Abraham et al. 2009). Our findings from study 1 are consistent with these previous findings in that both beginning and advanced students improved in their understanding of natural selection after completing the tutorial version of the module.

We hypothesized there might be an interaction between the version of the module and the academic level of the students because more advanced students might be more likely to benefit from the more open-ended nature of the workbook version than beginners would. In study 2, however, we found no evidence to support an interaction between module version and level of student. Additionally, in study 2 “advanced” students outperformed “beginner” students on the post-test. In interpreting results from both studies, we note that our definition of “advanced” encompasses a potentially wide range of previous exposure to natural selection, limiting the scope of the conclusions that can be drawn.

Limitations of approach

We used a design-based research approach, focusing on timely cycles of revision and assessment. Our approach has several limitations. First, the classes were recruited by convenience and not by any research-driven criteria. Second, there were weaknesses with the summative end-of-unit assessment used in study 2 to compare the two module versions. An assessment is all about the inferences one wants to make (Messick 1989) and the end-of-unit assessment was designed to allow instructors to infer whether or not students had completed all of the tasks in the module and focused on the content rather than just clicking through. The assessment contained only 10 items, and while it was based in part on the assessment used in study 1, some items were different and there were fewer items overall. Also, we were not able to collect pertinent information such as demographic data or prior knowledge (e.g., through a pre-test). A more rigorous study that randomly assigns students to treatment and uses the assessment from study 1 would better compare the effects of the two versions on students learning of natural selection.

Our revisions were evidence-based, and the immediate feedback made possible by the conversion to the tutorial style format with forced-response items was crafted to target misconceptions and help students reflect on their thinking. While the results demonstrate that student learning was not impacted when the module was converted from the workbook to the tutorial format, we cannot separate the effects of the format change itself from the many other substantive revisions to the module content we made as part of the iterative design-based research process (see Additional file 1: Tables S1 and S2).

Conclusion

Revising the *Darwinian Snails* module using a design-based research approach and adopting a feedback-intensive tutorial style allowed us to focus on the key concepts around natural selection and identify the misconceptions that students hold. Our studies show that this approach

allowed us to produce a module that is effective at teaching some important aspects of evolution by natural selection. Students showed learning gains on all targeted key concepts, and reduced expression of all targeted misconceptions, which was not found previously for students using the older workbook version of the module. In particular, following the design changes made to the tutorial version, students appeared to overcome the Adaptive mutation misconception; evidence for this had been lacking for the previous workbook version. Our iterative design-test-redesign approach while the module was being used in real classrooms, and our use of a different assessment, limits our ability to directly compare our results to the previous study.

More broadly, this study provides a strong example of successfully using a design-based research approach to guide improvements to established teaching tools.

Additional file

Additional file 1: Table S1. Significant changes between workbook and tutorial versions of the Darwinian Snails module as used in study 1. **Table S2.** Significant changes between initial revisions of tutorial version (used in study 1) and later revisions of tutorial version (used in study 2) of the Darwinian Snails module. **Table S3.** Breakdown of classes in study 1. **Table S4.** Classification of items used for pre/post assessment in study 1. **Table S5.** Sample student response to the ACORNS question. **Table S6.** Model taxonomy of hierarchical linear model results of pre and post-test data in study 1. **Table S7.** Summary of means for advanced and beginner classes in study 1. **Table S8.** Percent of students who either did or did not choose a different response on the pre-test and post-test. **Table S9.** Key concepts and misconceptions scored by EvoGrader that were not included in our key concepts and misconceptions. **Table S10.** Classes in study 2. **Table S11.** Classification of items used for pre/post assessment in study 2. **Table S12.** Model taxonomy of hierarchical linear model results of performance measure outcomes for study 2. **Table S13.** Distribution of responses chosen by students in the workbook or tutorial classes for question assessing the mutations random key concept on the summative assessment in study 2.

Abbreviations

AAAS: American Association for the Advancement of Science; ACORNS: Assessment of Contextual Reasoning about Natural Selection; COUHES: Committee on the Use of Humans as Experimental Subjects; DBR: design-based research; HLM: hierarchical or multilevel model; ICC: intraclass correlation; KC: key concept; MIT: Massachusetts Institute of Technology; NEIRB: New England Independent Review Board; SD: standard deviation; SE: standard error.

Authors' contributions

DSP and JCM conducted data analysis. All authors wrote and edited the manuscript. All authors read and approved the final manuscript.

Author details

¹ Department of Instructional Technology and Learning Sciences, Utah State University, 2830 Old Main Hill, Logan, UT 84321, USA. ² Center for the Integration of Research, Teaching and Learning, 1025 W. Johnson St, Madison, WI 53706, USA. ³ SimBio, 1280 S 3rd St W, Suite 3, Missoula, MT 59801, USA. ⁴ Department of Biological Science, California State University, Fullerton, CA 92834-6850, USA.

Acknowledgements

Thanks to Kerry Kim for help with data extraction. Thanks to the many faculty and students who contributed data to this project.

Competing interests

EM and SM are employees of SimBio. JKA, DSP and JCM are collaborating with SimBio on several projects.

Ethical approval and consent to participate

Study 1 was approved by the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology in Cambridge, MA before data collection (COUHES #1206005102), and for each of the classes whose data we used, we also received approval from the IRBs of their institutions (they either chose to review and approve the study or accepted the approval of MIT COUHES). Study 2 was approved by the New England Independent Review Board (NEIRB #120160152) to use the de-identified answer data for research purposes after the data had been collected.

Funding

This material is based upon work supported in part by the National Science Foundation under Grant No. 1227245.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 27 August 2017 Accepted: 10 April 2018

Published online: 26 April 2018

References

- Abraham JK, Meir E, Perry J, Herron JC, Maruca S, Stal D. Addressing undergraduate student misconceptions about natural selection with an interactive simulated laboratory. *Evol Educ Outreach*. 2009;2(3):393–404. <https://doi.org/10.1007/s12052-009-0142-3>.
- American Association for the Advancement of Science. AAAS science assessment website. Washington: American Association for the Advancement of Science; 2013. <http://assessment.aaas.org/>. Accessed 17 Apr 2018.
- Anderson DL, Fisher KM, Norman GJ. Development and evaluation of the conceptual inventory of natural selection. *J Res Sci Teach*. 2002;39(10):952–78. <https://doi.org/10.1002/tea.10053>.
- Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1–48. <https://doi.org/10.18637/jss.v067.i01>.
- Bishop BA, Anderson CW. Student conceptions of natural selection and its role in evolution. *J Res Sci Teach*. 1990;27(5):415–27. <https://doi.org/10.1002/tea.3660270503>.
- Bishop BA, Anderson CW. Evolution by natural selection: a teaching module. Occasional Paper No. 91. 1986.
- Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Erlbaum Associates; 1988.
- Collective, D. B. R. Design-Based Research Collective. Design-Based Research: an emerging paradigm for educational inquiry. *Educ Res*. 2003;1(32):5–8.
- Crocker L, Algina J. Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, 32887. 1986.
- DeVellis R. Scale development theory and applications. Thousand Oaks: SAGE Publications; 2003.
- Easterday M, Rees Lewis D, Gerber E. Design-based research process: problems, phases, and applications. In: Proceeding of international conference of learning sciences (vol 14). 2014.
- Ferrari M, Chi MT. The nature of naive explanations of natural selection. *Int J Sci Educ*. 1998;20(10):1231–56.
- Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP. Active learning increases student performance in science, engineering, and mathematics. *Proc Natl Acad Sci*. 2014;111(23):8410–5. <https://doi.org/10.1073/pnas.1319030111>.
- Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. *J Exp Psychol Gen*. 2012;141(1):2–18. <https://doi.org/10.1037/A0026092>.
- Garvin-Doxas K, Klymkowsky MW. Understanding randomness and its impact on student learning: lessons learned from building the Biology Concept Inventory (BCI). *CBE Life Sci Educ*. 2008;7(2):227–33.
- Gregory TR. Understanding natural selection: essential concepts and common misconceptions. *Evol Educ Outreach*. 2009;2(2):156–75. <https://doi.org/10.1007/s12052-009-0128-1>.
- Hake RR. Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. *Am J Phys*. 1998;66(1):64–74.
- Herron JC, Maruca S, Meir E. Darwinian Snails. Missoula: SimBio; 2014. <http://www.simbio.com>. Accessed 17 Apr 2018.
- Jensen MS, Finley FN. Changes in students' understanding of evolution resulting from different curricular and instructional strategies. *J Res Sci Teach*. 1996;33(8):879–900.
- Kalinowski ST, Leonard MJ, Andrews TM, Litt AR. Six classroom exercises to teach natural selection to undergraduate biology students. *CBE Life Sci Educ*. 2013;12(3):483–93. <https://doi.org/10.1187/cbe-12-06-0070>.
- McNemar Quinn. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947;12(2):153–7.
- Meir E, Perry J, Herron JC, Kingsolver J. College students' misconceptions about evolutionary trees. *Am Biol Teach*. 2007;69(7):e71–6.
- Messick S. Meaning and values in test validation: the science and ethics of assessment. *Educ Res*. 1989;18(2):5–11.
- Mohareri K, Ha M, Nehm RH. EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evol Educ Outreach*. 2014;7(1):15.
- Nehm RH. Faith-based evolution education? *Bioscience*. 2006;56(8):638–9.
- Nehm RH, Reilly L. Biology majors' knowledge and misconceptions of natural selection. *Bioscience*. 2007;57(3):263–72.
- Nehm RH, Schonfeld IS. Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. *J Res Sci Teach*. 2008;45(10):1131.
- Nehm RH, Beggrow EP, Opfer JE, Ha M. Reasoning about natural selection: diagnosing contextual competency using the ACORNs instrument. *Am Biol Teach*. 2012;74(2):92–8.
- Pope DS, Rounds CM, Clarke-Midura J. Testing the effectiveness of two natural selection simulations in the context of a large-enrollment undergraduate laboratory class. *Evol Educ Outreach*. 2017;10:3. <https://doi.org/10.1186/s12052-017-0067-1>.
- Price RM, Pope DS, Abraham JK, Maruca S, Meir E. Observing populations and testing predictions about genetic drift in a computer simulation improves college students' conceptual understanding. *Evol Educ Outreach*. 2016;9:8. <https://doi.org/10.1186/s12052-016-0059-6>.
- R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2017. <http://www.R-project.org/>. Accessed 17 Apr 2018.
- Resnick LB, Resnick DP. Assessing the thinking curriculum: new tools for educational reform. In: Gifford BR, O'Connor MC, editors. Changing assessments. Evaluation in Education and Human Services, vol. 30. Dordrecht: Springer; 1992. p. 37–75. https://doi.org/10.1007/978-94-011-2968-8_3.
- Robbins JR, Roy P. The natural selection: identifying and correcting non-science student preconceptions through an inquiry-based, critical approach to evolution. *Am Biol Teach*. 2007;69(8):460–6.
- Rosenthal R. Parametric measures of effect size. In: Cooper H, Hedges LV, editors. The handbook of research synthesis. New York: Russell Sage Foundation; 1994. p. 231–44.
- Seeley RH. Intense natural selection caused a rapid morphological transition in a living marine snail. *Proc Natl Acad Sci*. 1986;83(18):6897–901.
- Smetana LK, Bell RL. Computer simulations to support science instruction and learning: a critical review of the literature. *Int J Sci Educ*. 2012;34(9):1337–70. <https://doi.org/10.1080/09500693.2011.605182>.
- Soderberg P, Price F. An examination of problem-based teaching and learning in population genetics and evolution using EVOLVE, a computer simulation. *Int J Sci Educ*. 2003;25(1):35–55. <https://doi.org/10.1080/09500690110095285>.
- Theobald R, Freeman S. Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. *CBE Life Sci Educ*. 2014;13(1):41–8. <https://doi.org/10.1187/cbe-13-07-0136>.
- Quellmalz ES, Pellegrino JW. Technology and testing. *Science*. 2009;323(5910):75–9.