Pope *et al. Evo Edu Outreach (2017) 10:3*
DOI 10.1186/s12052-017-0067-1

Evolution: Education and Outreach

RESEARCH ARTICLE Open Access

Testing the effectiveness of two natural selection simulations in the context of a large-enrollment undergraduate laboratory class

Denise S. Pope[1,4*], Caleb M. Rounds[2] and Jody Clarke-Midura[3]

Abstract

Background: Simulations can be an active and engaging way for students to learn about natural selection, and many have been developed, including both physical and virtual simulations. In this study we assessed the student experience of, and learning from, two natural selection simulations, one physical and one virtual, in a large enrollment introductory biology lab course. We assigned students to treatments (the physical or virtual simulation activity) by section and assessed their understanding of natural selection using a multiple-choice pre-/post-test and short-answer responses on a post-lab assignment. We assessed student experience of the activities through structured observations and an affective survey.

Results: Students in both treatments showed increased understanding of natural selection after completing the simulation activity, but there were no differences between treatments in learning gains on the pre-/post-test, or in the prevalence of concepts and misconceptions in written answers. On a survey of self-reported enjoyment they rated the physical activity significantly higher than the virtual activity. In classroom observations of student behavior, we found significant differences in the distribution of behaviors between treatments, including a higher frequency of off-task behavior during the physical activity.

Conclusions: Our results suggest that both simulations are valuable active learning tools to aid students' understanding of natural selection, so decisions about which simulation to use in a given class, and how to best implement it, can be motivated by contextual factors.

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

Background

An understanding of evolution by natural selection is fundamental to biological literacy and yet is a notoriously difficult topic for students, with misconceptions that persist even after instruction (Bishop and Anderson 1990; Nehm and Reilly 2007; Nehm and Schonfeld 2008; Gregory 2009). In order to help students grapple with and overcome these misconceptions, we need approaches that employ active learning (Alters and Nelson 2002; Nelson 2008) and are proven effective (see critique by Nehm 2006). Since natural selection is difficult to observe directly in the context of a class (but see Krist and Showsh 2007; Plunkett and Yampolsky 2010; Serafini and Matthews 2009), many educators have turned to active learning approaches that use simulations of the process to dispel student misconceptions.

Simulations of natural phenomena are useful teaching tools because they allow visualization of temporal and spatial scales that are either too large or too small to be directly accessed by students and enable students to investigate the underlying factors that influence these phenomena by changing variables and observing the

*Correspondence: denise.pope@wisc.edu
[4] Present Address: Center for the Integration of Research, Teaching and Learning, 1025 W. Johnson St., Madison, WI 53706, USA
Full list of author information is available at the end of the article

outcome (National Research Council 2011; Rutten et al. 2012; Smetana and Bell 2012). Natural selection simulations can help dispel misconceptions about the process because students can see, as they view the simulation, that individuals do not change but instead that differential survival and reproduction combined with heritability change the composition of the next generation, and that selection acts on pre-existing variation rather than inducing variation. Simulations that have been developed for teaching natural selection have included both simulations in which students physically manipulate objects or their own body to represent the population (Fifield and Fall 1992; Van Thiel 1994; Siegel et al. 2005; Price 2011; Eterovic and Santos 2013; Hildebrand et al. 2014), and also simulations of virtual populations (Latham and Scully 2008; Abraham et al. 2009; BraySpeth et al. 2009; Soderberg and Price 2003; Yamanoi and Iwasaki 2015).

Like many proposed interventions to teach natural selection (Nehm 2006), the effectiveness of many simulations (both physical and virtual) designed to teach natural selection has not been tested. Therefore, we designed this study to assess the effectiveness of two natural selection simulations, one physical and one virtual, in the context of a college introductory biology laboratory course. Both simulations are widely used in college biology classes in the United States. In addition to assessing each activity, we also compare the two activities to investigate any differences in learning gains, student experience, and details of implementation.

Physical vs. virtual learning activities

One of the most salient differences between the two activities we tested in our study is the physical or virtual nature of the simulation. While this difference might be expected to impact student learning, previous research does not support the idea that there is an inherent advantage of interaction with either physical or virtual materials. Many studies in K-12 STEM education have compared physical and virtual activities that teach the same concepts, and most have found no difference in learning (Chen et al. 2014; Klahr et al. 2007; Lazonder and Ehrenhard 2014; Marshall et al. 2010; Renken and Nunez 2013). Studies at the undergraduate level (mostly in STEM fields outside of biology) generally compare lab activities using physical equipment to activities with virtual lab equipment and have found similar trends. While a few college-level studies have shown some advantage to either the physical (Taghavi and Colen 2009) or the virtual (Pyatt and Sims 2007) versions of an activity, most studies found no differences in learning gains (Corter et al. 2007; Hawkins and Phelps 2013; Kelly et al. 2008; Nickerson et al. 2007; Tatli and Ayas 2013).

Student learning is not the only dimension of the student experience, and since most studies comparing virtual and physical activities show no difference in learning, focusing on other aspects of student experience can better elucidate how the activities differ and in what context one of them might be preferable. Some of these studies have surveyed students about their preference for or attitudes about the physical and virtual activities (Corter et al. 2007; Dewhurst et al. 1994; Pyatt and Sims 2007; Scoville and Buskirk 2007) or asked them to self-assess their learning gains (Wiesner and Lan 2004). A few studies investigated the behavior of students during the activities, using informal observations (Finkelstein et al. 2005; Steinberg 2000) or gaze tracking technology (Chien et al. 2015).

In most studies comparing virtual and physical activities, the virtual activity is a simulation of the physical activity; i.e., they simulate the human experience of using physical lab equipment or moving around in the physical lab environment. In contrast, both the physical and the virtual activities in our study are simulations of a biological process; neither activity allows direct perception of the phenomenon being studied. In this sense, neither of the activities we test is more "real" than the other, so there is no reason to posit an advantage to hands-on manipulation for learning the concepts of evolution by natural selection.

Testing natural selection simulations

In this study, we implement one physical and one virtual simulation of natural selection in a large-enrollment laboratory class, we assess student learning and self-reported enjoyment, and conduct structured observations of student behavior during the activities. Our aims are: first, to assess the effectiveness of each of the simulations in support of evidence-based teaching (Nehm 2006), and second, to compare the two simulations—to our knowledge, natural selection simulations have not been directly compared.

We address the following research questions:

1. Do each of the simulations (physical and virtual) significantly increase student understanding of key concepts of natural selection?
2. Do measures of student learning significantly differ between the two simulations?
3. Do measures of student engagement significantly differ between the two simulations?

Methods

Lab activities

Physical natural selection simulation (Clipbirds)

The Clipbirds simulation is based on a procedure developed by Janulaw and Scotchmoor (2011). Briefly, during three timed "seasons" students act as birds competing for food resources in two different environments. Birds that

collect enough food survive, and if very successful, reproduce; unsuccessful birds die. Over the course of four seasons, students observe how phenotype frequencies shift in the population. Bird beaks are represented by three sizes of binder clips. Food resources are represented by three sizes of round objects: marbles, garbanzo beans and popcorn kernels. After each season, students place alleles of survivors into a bag, which allows for random assortment. Two tables simulate two islands that have slightly different food resources each season (for full procedures, see Additional file 1). A second exercise works the same way but adds in genetic drift, with a dice roll after the first season that wipes out all of the birds of one phenotype.

Virtual natural selection simulation (Darwinian Snails)

The SimBio Virtual Labs® module Darwinian Snails (Herron et al. 2014) uses as a framework Robin Seeley's study on the effect of European green crab predation on the evolution of shell thickness in periwinkle snails in New England (Seeley 1986), and the module is structured around a series of simulations of a periwinkle snail population. In the first part of the module the students play the role of crabs eating snails, and are told to maximize their efficiency while doing so, which usually leads them to focus on snails with the thinnest shells. They can then see how this affects the distribution of shell thicknesses over the next three generations. After that introduction, students add crabs to the snail population and set parameters and control the simulation, but are no longer participating in the simulation. They are asked to run the simulation sequentially without trait variation, without trait heritability, and without differential survival, to demonstrate that each are necessary conditions for natural selection. They are introduced to mutation as the generator of the variation present in the population, and they observe that the direction of mutations is random (offspring may have shells that are either thicker or thinner than their parents), and that the presence of crabs does not affect the direction of mutations. The module includes questions and guiding text throughout that is designed both to increase students' understanding of the conditions of natural selection, and also to confront common misconceptions about natural selection. An earlier version of the simulation is described in detail in Abraham et al. (2009). The version used in our study has all of the instructions onscreen and includes questions throughout the activity (mostly multiple-choice) with immediate feedback (Clarke-Midura et al. unpublished observations).

Selection of and comparison of the two lab activities

We chose the two activities based on practical considerations. The physical simulation had already been in use in the course for several years. The instructor was exploring alternative activities and was interested in trying the virtual simulation but wanted to be sure that it was equally effective as the physical simulation. Since neither activity had been assessed in the context of a large-enrollment laboratory class, we used the opportunity to design a study to accomplish this assessment and to compare the two activities.

The physical and virtual simulations both rely on the central theme of predator–prey interactions leading to selection for adaptive traits, but there are some significant differences. In the physical exercise, students calculate the reproductive success of the birds—plastic chips represent the alleles of each bird, and after each generation the birds reproduce based on the frequency of these alleles in a bag, thus simulating recombination. In the virtual simulation, the software computes reproductive outcomes and while heritability is discussed, there is no explanation of the genetic basis of the selected trait, and snails reproduce by cloning. The physical simulation also includes the concept of genetic drift in some of the simulated generations while the virtual one discusses drift only briefly (as evolution in the absence of differential survival). The questions with immediate feedback that are included in the virtual simulation are absent in the physical simulation. The physical simulation is participatory—students act as members of the evolving population, whereas for most of the virtual exercise, students set parameters and initiate the simulation but are not actors in the simulated population. These differences, as well as the physical vs. virtual distinction, might impact learning gains and student experience.

Study design and implementation

Description of sample and assignment to treatments

This study was conducted at a public university in the northeast United States, and the use of human subjects was reviewed and approved by the university's Institutional Review Board (protocol# 2013-1885). We carried out the study in the context of a stand-alone laboratory class that is required of all biology majors, with an enrollment of over 800 students. In order to enroll in the laboratory class, students must have passed the first semester biology course (focused on cellular and molecular biology) with a grade of C or better. Most students in the lab class were simultaneously enrolled in the second semester biology course, which is focused on organismal biology, ecology, and evolution.

Students were enrolled in 36 sections, with an average of 23.5 $\pm$ 3.6 students per section, taught by 20 teaching assistants (TAs). Most of the TAs taught two sections. In order to prevent the TAs from having to prepare to teach both lab exercises, we randomly assigned sections to

treatment at the level of the TA, balancing for time of day and day of the week.

The baseline criterion for inclusion in the study was that a student completed the pre-test and consented to participate. Most students (over 90%) who took the pre-test consented to participate, for a total of 640 consenting students (~77% of the class). This includes 307 in the physical treatment and 333 in the virtual treatment. Our final sample for each analysis is a subset of these 640, depending on how many of those students completed the other assessments used in a given analysis.

About 2 weeks after students had completed the lab, along with an enjoyment survey (see below), we administered an optional short demographic survey that asked students their gender and whether or not English was their native language. 451 consenting students completed the survey (~55% of the class); of these, 65% of students self-identified as female and 16% reported that English was not their native language.

Implementation

All assessments and assignments used in this study were implemented through the University's online course management system, Moodle. All students were assigned a pre-lab assignment that consisted of a reading assignment from the textbook and a pre-lab quiz on the reading. Students had to complete the quiz by midnight the night before their lab. The pre-lab reading quiz was automatically graded by Moodle and was assessed as part of the student's grade for the semester. Students were also given the optional assignment to take an assessment on their understanding of natural selection (pre-test) by midnight the night before their lab. The informed consent form preceded the pre-test.

During the lab, students either completed the Clipbirds lab (physical treatment) or Darwinian Snails lab (virtual treatment) described above. After completing the lab, students were assigned a post-lab assignment consisting of multiple-choice and short-answer questions (see Additional file 1), which was graded by their TA. Students had the text of these questions available during lab and were asked to discuss them with their group, but they then had to write their own answers and submit them through Moodle. They were also given the option of answering the natural selection multiple-choice assessment (post-test) and a survey that asked them to rate their enjoyment of the lab activity. Four weeks after completing the lab, students were again given the option of completing the natural selection assessment (delayed post-test).

The natural selection assessments (pre-, post- and delayed post-tests) and the survey were not required for course credit, and students never learned their score on the assessments. Students were given the option to take these for extra credit (1 point for the pre- and post-tests and the survey and 5 for the delayed post-test). The extra credit amounted to less than 1% of their final grade. Students who chose not to participate in the study (i.e. did not provide consent) but completed the assessments were still awarded extra credit. The instructor and TAs were not aware of which students consented to participate in the study. Because completion of the pre- and post-tests and the survey earned extra credit, our sample may be biased towards more highly motivated students, but given the high number of students who completed those assessments, we nonetheless have data from more than half of the class for each of those analyses.

Outcome measures

Natural selection assessment

In order to measure student understanding of natural selection concepts, we used a 14-item content test we had previously compiled (Clarke-Midura et al. unpublished observations), based on items in three different published instruments: AAAS Science Assessment Website (2013), the Concept Inventory on Natural Selection (Anderson et al. 2002), and the Natural Selection Diagnostic Test (Bishop and Anderson 1990). We selected these 14 multiple-choice items to specifically target the key concepts we had identified (Clarke-Midura et al. unpublished observations). These focused on students' understanding of the basic requirements for natural selection: differential survival and reproduction based on heritable traits, variation of these traits in the population, and random mutations as a source of new traits. The test also addresses a common misconception that organisms change their traits because they "need" to (see Additional file 1 for full text of assessment). We removed one item (question 3) from the analysis due to a typo in the answer selection on two of the tests. Therefore, we used only 13 items in our analysis. The item we removed was the most difficult one on the test (Clarke-Midura et al. unpublished observations), making the average level of performance higher when including only the 13 items. Cronbach's alpha, a measure of internal consistency, for the pre-, post-, and delayed post-test were 0.78, 0.80, and 0.82 respectively.

We present the scores for the 13-items on a 100-point scale for ease of comparison. 481 students (~58% of the class) consented to participate and completed all three natural selection assessments (pre-test, post-test and delayed post-test): 226 students in the physical treatment and 255 students in the virtual treatment. To estimate the effect size of the increase from pre-test to post-test and pre-test to delayed post-test within each treatment, we calculated Cohen's d [(post-test mean − pre-test mean)/(pooled SD)] (Cohen 1988).

To assess student learning and compare learning between the two treatments, we conducted two analyses. First, we calculated normalized change scores between the pre-test to the post-test, and the pre-test to the delayed post-test. This method was developed by Marx and Cummings (2007) as a student-level alternative to normalized gain scores that are typically done at the classroom level (Hake 1998; Theobald and Freeman 2014). We tested the assumption of normality for the normalized gain scores with a Shapiro–Wilk's W test; since normality was violated we used non-parametric Wilcoxon–Mann–Whitney rank-sum tests to compare treatments. When calculating normalized change scores, students who score 100 on both tests are dropped from the sample, so the sample size for this is smaller (physical n = 211 for both pre to post and pre to delayed post; virtual n = 234 for pre to post, 239 for pre to delayed post comparison).

Second, we conducted a repeated-measures ANOVA to test for differences in students performance at the three different time points (pre, post, and delayed post). In our model, time was our within-subject factor and treatment, gender (male or female), and language (native English speaker or non-native speaker) were the between-subject factors. We tested the main effect (time) and its interactions with treatment, gender, and language, and used Wilcoxon signed-rank tests to make post hoc comparisons. Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2_{(2)} = 9.455$, $p < 0.01$), therefore we corrected degrees of freedom using Huyn and Feldt estimates of sphericity ($\varepsilon = 0.999$). Only students who completed all three tests and the demographic survey were included in this analysis; we also included only students who self-identified as male or female, since the number of students who identified as other genders were too small to include as a factor in the analysis. The total for this analysis was 419 students (200 in physical treatment; 219 in virtual treatment), representing about 50% of the students in the class.

Post-lab assignment

The post-lab assignment (Additional file 1) consisted of two short-answer questions that were specific to each lab activity (physical or virtual) and six questions that were the same for all students and can therefore be used to compare student responses between treatments. The six questions consisted of three pairs of multiple-choice and short-answer questions; one pair of questions was designed specifically to assess students' understanding of genetic drift. Student responses were graded by their TA, but for the purposes of this study, we coded a random subset of the responses of consenting students (81 students from the physical treatment and 89 from the virtual treatment).

We coded student responses to the short-answer questions (blinded to treatment) for the presence of concepts relating to natural selection and genetic drift, and the presence of evolutionary misconceptions. Two researchers read through a subset of answers to determine which concepts and misconceptions occurred in student responses for each question and then consulted to develop codes used for scoring. The final codes for the short-answer questions included six concepts and four misconceptions.

Both researchers then coded 20 student responses. The concept and misconception codes were not mutually exclusive, each were only coded when present, and some of the codes occurred only rarely. In this situation, if we counted as "agreement" for a given code all cases where we both did not use a code, in addition to those where we did use a code, we would overestimate agreement. Therefore, to compare inter-rater reliability, we calculated the Jaccard coefficient for each code. The Jaccard coefficient corrects for "negative agreement" by not including the number of instances where a code was assigned by neither coder, and is used as a measure of similarity in various contexts (Cheetham and Hazel 1969; Real and Vargas 1996; Smith et al. 2013). Like other measures of inter-rater agreement, the Jaccard coefficient ranges between 0 (no agreement)—1 (complete agreement).

The Jaccard coefficient equation is

$$J = \frac{A}{A + B + C}$$

where A = # of instances code was assigned by both coders, B = # of instances code was assigned only by 1st coder, C = # of instances code was assigned only by 2nd coder.

The average Jaccard coefficient for all 10 codes was 0.67 ± 0.19. As another measure of inter-rater reliability, we also calculated Cohen's κ, which corrects for the probability of two raters agreeing due to chance, but includes both negative agreement and positive agreement. The mean κ was 0.72 ± 0.14. After reaching consensus on the 20 responses we both coded, one of us coded an additional 170 student responses, which are the ones used in our analysis.

We calculated the frequency of concept and misconception use by students in the two treatments. Some of the concepts and misconceptions occurred in responses to more than one of the questions; for these, we counted them as "present" for a student if they occurred in at least one item, so we did not count any concept or misconception twice for a student who used it twice. We compared mean frequencies of concept and misconception use using nonparametric Wilcoxon–Mann–Whitney rank-sum tests but found no significant differences

between treatments, even without correcting for multiple comparisons.

Engagement measures

Student attitude survey

In order to assess the students' enjoyment of the lab, we included a short Likert-scale survey two weeks after they completed the lab. Students responded to questions such as "I would be willing to do this lab activity again because I think it was fun" on a scale from 1 to 6 where 1 = Completely false to 6 = Completely true. The survey also included other items about student self-efficacy on the survey, but we did not include these in the analysis because there were no differences between treatments in the self-efficacy scores, and they were not relevant to our research questions. The items were from the Enjoyment/Interest subscale of the Intrinsic Motivation Inventory (Deci et al. 1994) and were modified for the context of the labs (see Additional file 1). The internal consistency (Cronbach's alpha) for the four items asking about enjoyment/interest was 0.93. The scores were averaged across the four items to calculate each student's composite "enjoyment rating." The sample size of consenting students who completed this enjoyment survey was 216 in the physical treatment and 235 in the virtual treatment (~53% of the class). Because these are ordinal data, we conducted a nonparametric Wilcoxon–Mann–Whitney rank-sum test to compare enjoyment rating by treatment.

Classroom observations

In addition to student self-reporting, we wanted to assess how students engaged with each of the lab activities through direct observation of students in lab classrooms. Many classroom observations methods score student behavior in aggregate (Eddy et al. 2015; Sawada et al. 2002; Smith et al. 2013), so the class as a whole is scored as participating in an activity. That method is appropriate for providing teachers with feedback on their classroom but ignores variation among students in levels of engagement, and so is less appropriate for our purposes. We wanted to observe each student in a lab section to get a representative sample and to assess the variation in student involvement in the lab activities.

We developed an observation protocol that focused on observable behaviors of individual students. We developed the method because we were not able find an existing protocol that fulfilled our criteria that it: focused on students not instructors, observed individual students, did not require subjective estimations of engagement, and allowed us to capture the range of behaviors we were interested in and that we could easily observe from several feet away. Because of the structure of lab classrooms, where students generally remain in the same general location and belong to a lab group that is stable (at least within one lab period), it was feasible to repeatedly observe the same individuals. In developing our observation protocol, we confronted two issues: how to define engagement, and how to observe and measure it.

For the purposes of this study, we defined engagement as including both the level of involvement in an activity (sometimes called behavioral engagement), and the interest in/enjoyment of that activity (cognitive and emotional engagement). In this sense, our use of the term engagement is a composite state involving both internal and external aspects. The involvement aspect is observable, while we can only observe correlates of interest and enjoyment, such as smiling, laughing, or a look of concentration. We chose observable behaviors to minimize the potential for subjective judgments of students as being "engaged" or "not engaged". We recorded behavior in four categories: (1) the direction of gaze, (2) verbal interactions, (3) motor behavior (primarily hands/arms), (4) facial expression/affect.

We first specified comprehensive and mutually exclusive behaviors to record within each category, and after two practice rounds of observations, we discussed and refined the list of behaviors and agreed on a definition of each behavior. The verbal and affect categories had four behaviors, the gaze category six, and the motor category eight for the physical activity or seven for the virtual activity (Table 1).

Before starting observations in each classroom, we identified each lab group (2 groups in physical treatment and 4–8 groups in virtual treatment) and the number of students in each group (11–13 students in the physical treatment and 2–4 students in the virtual treatment). We haphazardly assigned numbers to groups and to students within the group (e.g. Group 1, Student 1) and then recorded their (presumed) gender and identifiable clothing that they were wearing (e.g., red hat) so that we could identify them in multiple rounds of observation. Therefore, when we started observations and recording behavior, the lab activities were underway.

We employed scan sampling (developed by Altmann 1974) for the observations; sequentially observing individual students. Starting with a group at one end of the room, we observed Student 1 in that group and recorded one behavior of in each of the four behavior categories (gaze, verbal, motor, and affect) for that student, and then moved on to the next student in the group. Observations of each student took less than 5 s. After making one observation of each student in each group in the class, we paused before starting the next round of observations, which we conducted in the same order as the initial round. Therefore, each student was observed several times and the time elapsed between observations was

Table 1 Four behavioral categories and mutually exclusive and comprehensive behaviors recorded in each category

	Gaze	Verbal	Motor		Affect
			Physical	Virtual	
Engaged	Activity	Peer	Activity-related	Activity-related	Focus
	Paper	TA	Using lab objects	Mouse	Positive
	Peer		Typing	Typing	
	TA		Pointing	Pointing	
			Walking		
			Writing	Writing	
Neutral		No verbal	No motor	No motor	
Unengaged	Phone	Peer off-task	Texting	Texting	Negative
	Other off-task		Other off-task	Other off-task	Focus off-task

Most behaviors in each category suggest students are engaged or unengaged in the simulation activity, but verbal and motor categories include the possibility of no behavior, which cannot be classified as either engaged or unengaged. Motor categories were differentiated by treatment due to the different nature of the physical and virtual activities

roughly equivalent for all students in a section. Scan sampling is a useful technique to estimate the behavioral distribution of individuals in groups (Altmann 1974), and is frequently used in studies of animal behavior.

While students were aware of our presence in the room, we stayed at the periphery of the room and attempted to remain inconspicuous, and most students were focused on the lab activity and did not appear distracted by our presence.

Two researchers conducted the observations. After practicing on two class sections and determining our final behaviors in each category (see above), the two of us observed the same two class sections simultaneously (one in each treatment), ensuring that we were observing the same students at the same time by nodding or other nonverbal communication. We calculated inter-rater reliability of our observations for these two jointly observed sections using Cohen's κ (Table 2), which corrects for the probability of two observers agreeing by chance. Since the two treatments necessarily included different behaviors (Table 1), we calculated agreement separately for each treatment. Cohen's κ was between 0.5 and 0.8 for all categories, suggesting moderate to good agreement. The Affect category showed the most consistent inter-rater reliability across the two treatments.

After those jointly observed sections that we used to calculate inter-rater reliability, we individually observed 10 lab sections (5 of each treatment) over the course of 4 days, balancing across time of day. These are the observations we include in our analysis. In those 10 sections, we observed a total of 120 students in the physical treatment and 110 in the virtual treatment; this represents ~28% of all students in the course.

We recorded the behavior of each student in the section about 5 times (average of 4.9 for the physical treatment and 4.8 for the virtual treatment). Because of the different nature of the lab activities, group size differed between treatments: students in the physical sections were divided into two groups, with an average of 12.0 students per group, while students in the virtual sections worked in smaller groups (average 3.3), and there were on average 7 groups per section.

Table 2 Inter-rater reliability (Cohen's κ) and percent agreement for student observations

Behavior category	Physical treatment (36 observations)		Virtual treatment (60 observations)	
	% Agreement	Cohen's κ	% Agreement	Cohen's κ
Gaze	0.72	0.50	0.97	0.77
Verbal	0.83	0.61	0.80	0.52
Motor	0.75	0.65	0.88	0.79
Affect	0.86	0.76	0.93	0.75

By two observers in two class sections (one for each treatment)

For analysis of the observation data, we collapsed some of the subcategories in each behavioral category: for the Motor category, we lumped all of the activity-related behaviors; for the Gaze and Motor categories, we lumped all off-task activities. We then calculated the proportion of observations for each treatment in which we recorded each of the mutually exclusive behaviors (# of times behavior observed/total # observations). This analysis includes all observations so the sample is number of observations (physical: 590, virtual: 547), not number of students. We used χ^2 tests of independence to compare the distribution of behaviors in each category between the two treatments. As an estimate of effect size, we calculated Cramér's V, which is a measure of the strength of association for contingency tests (Cramér 1946; Siegel and Castellan 1988). Because we intentionally did not

know the identity of the students we observed, we cannot compare observations of students in the classroom with their performance on the assessments.

Results

Students' performance in both treatments improved from pre-test to post-test

In both treatments, student performance on the natural selection assessment improved after completing their lab activity. The average pre-test scores were equivalent between treatments (Mean ± SD on a 100 point scale: physical 71.24 ± 22.27; virtual 71.55 ± 21.08). In the physical treatment, the average score on the post-test increased to 83.90 ± 19.5 with an effect size (Cohen's d) of 0.60; and on the delayed post-test it remained higher than the pre-test (78.46 ± 22.28; Cohen's d = 0.32). Similarly, in the virtual treatment, the average score on the post-test was greater (84.13 ± 18.52; Cohen's d = 0.63); and remained higher on the delayed post-test (77.47 ± 22.85; Cohen's d = 0.27). For both treatments, those changes represent medium effect sizes for the pre-test to post-test comparison, and small effect sizes for the pre-test to delayed post-test comparison, suggesting that both lab exercises were effective at teaching natural selection, and that the effect persisted for several weeks.

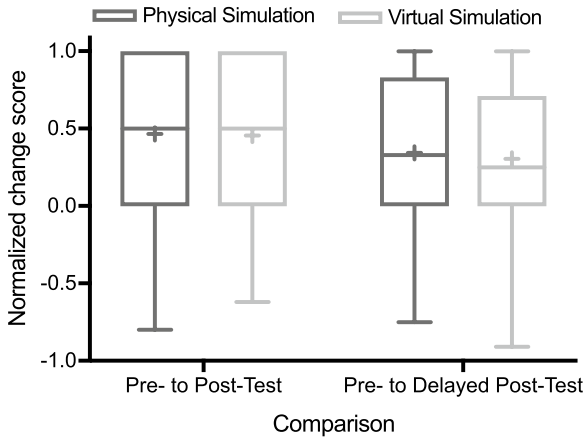

Fig. 1 Learning gains from the physical and virtual simulation activities. Students completed identical natural selection assessments before the lab, after the lab, and then 1 month after their lab. Normalized change represents the fraction of possible score improvement that the student achieved. *Centerlines* show the medians; *box limits* indicate the 25th and 75th percentiles; whiskers extend to the minimum and maximum values. Means are indicated by + symbol. Students in both treatments showed an improvement from pre- to post-test and pre- to delayed post-test. There were no significant differences in normalized gain scores between the treatments for either comparison

Students showed no difference in learning gains by treatment

To compare student performance between treatments, we calculated normalized change scores as a measure of student learning gains and conducted a repeated measures ANOVA. Normalized change is the fraction of potential improvement achieved by each student; 1 indicates improvement to a perfect score and 0 means no change from the earlier score.

The normalized change was essentially the same for both treatments (Fig. 1). Students in the physical (n = 211) treatment had a mean normalized change of 0.47 ± 0.46, with a median of 0.50, and those in the virtual treatment (n = 234) had a mean normalized change of 0.46 ± 0.47, with a median of 0.50. To evaluate how much students retained their knowledge of natural selection, we compared the normalized changed score between the pre-test to the delayed-post test. Students in the physical treatment had a mean normalized change of 0.34 ± 0.49 (median = 0.33) from pre-test to delayed post-test, and those in the virtual treatment had a normalized change of 0.30 ± 0.48 (median = 0.25).

There were no significant differences between the treatments in either comparison: pre- to post-test ($Z = 0.078$, $p > 0.1$, $r = 0.004$), or pre- to delayed post-test ($Z = 0.505$, $p > 0.1$, $r = 0.023$), where r (which equals $Z/\sqrt{n}$) is a measure of effect size. Thus we have no evidence that either activity elicited greater learning gains.

In the repeated-measures ANOVA (Table 3), the only significant factor in our model was time $F_{1.99, 821.23} = 30.25$, $p < 0.001$. Treatment, gender, or whether or not their native language was English were not significant predictors. We conducted Wilcoxon signed-rank tests to make post hoc comparisons between performance on each of the three occasions, regardless of treatment; there was a significant different from pre- to post-test ($Z = 8.43$, $p < 0.001$, $r = 0.56$, from pre- to delayed post-test ($Z = 5.67$, $p < 0.001$, $r = 0.04$) and from post- to delayed post-test ($Z = 3.47$, $p < 0.001$, $r = 0.23$).

Students showed no difference in post-lab written responses

To assess how often students use common misconceptions and key concepts when writing about natural selection in their own words, we compared responses to graded post-lab assignment consisting of short answers prompts and multiple-choice questions. There were no differences between treatments in scores on the three multiple-choice questions (mean scores: physical 2.37 ± 0.73, virtual 2.36 ± 0.68). More than 85% of students in both treatments were able to correctly answer the multiple choice question focused on natural selection and a true/false statement about a natural selection misconception, while

Table 3 Results of repeated-measures analysis of variance (ANOVA)

Effect	Sum of squares	*df*	Mean square	F	p
Time	12,259.83	1.99	6135.67	30.25	*<0.0001*
Time × treatment	489.55	1.99	245.00	1.21	0.30
Time × gender	95.80	1.99	47.95	0.24	0.79
Time × english	310.79	1.99	155.54	0.77	0.47
Time × treatment × gender	273.30	1.99	136.78	0.67	0.51
Time × treatment × english	308.83	1.99	154.56	0.76	0.47
Time × gender × english	463.92	1.99	232.18	1.15	0.32
Time × treatment × gender × english	172.12	1.99	86.14	0.43	0.65
Error	166,559.45	821.23	202.82		

Table 4 Occurrence of key concepts in student responses to three short-answer questions in post-lab assignment

Key concept	# of items	Frequency	
		Physical	Virtual
Variation in trait	2	0.69	0.64
Differential survival/reproduction	2	0.64	0.60
Heritability of trait	2	0.44	0.45
Mutation as source of variation	1	0.19	0.29
Randomness of mutation	1	0.11	0.09
Genetic drift	1	0.04	0.06

Some concepts appeared in responses to more than one item. There were no significant differences between treatments in the frequencies of any of the concepts (χ^2 tests, all $p > 0.1$)

Table 5 Occurrence of misconceptions in student responses to three short-answer questions in post-lab assignment

Misconception	# of items	Frequency	
		Physical	Virtual
Mutations are adaptive responses to environment	1	0.53	0.60
Traits change because of need	2	0.49	0.43
Traits of individuals change	1	0.21	0.17

One misconception appeared in responses to two items. There were no significant differences between treatments in the frequencies of any of the misconceptions (χ^2 tests, all $p > 0.1$)

only about 60% in each treatment were able to answer the multiple choice question about genetic drift correctly. For the short-answer questions, we compared the frequency of key concepts and misconceptions in student responses between the two treatments, calculated from the presence of the concept or misconception in their written response to short answer questions. The frequencies of both concepts (Table 4) and misconceptions (Table 5) were very similar in both the physical and virtual treatments and there were no significant differences in any of the comparisons ($p > 0.1$ in all χ^2 tests).

Variation and differential survival/reproduction, concepts that were emphasized as learning objectives in both the physical and virtual simulation activities, were the most common concepts appearing in short answer questions. Other concepts, such as the random nature of mutation, the requirement that a trait be heritable for natural selection to occur, and using genetic drift to explain some changes, were less frequent but still occurred in similar frequencies in student responses in both treatments. Two misconceptions were fairly common in student answers in both treatments: mutations are an adaptive response to the environment, and traits change because of need (Table 5). Most students in both treatments (56% in physical; 54% in virtual) included a combination of key concepts and misconceptions in their responses to short answer questions; only 25% of students in the physical treatment and 21% of those in the virtual treatment expressed solely key concepts in their answers. There was no significant difference between treatments in the proportion of students expressing only concepts, only misconceptions, or both ($\chi^2 = 0.7$, $p > 0.1$).

Students reported enjoying the physical simulation more

In order to assess student enjoyment with the simulations, we administered a short survey after they completed their assignments and post-tests. Students responded to questions about their interest in and enjoyment of the activity on a scale from 1 to 6 where 1 = Completely false to 6 = Completely true. Higher scores indicate higher levels of enjoyment. Students in the physical treatment reported a significantly higher enjoyment rating (Median = 5, "Mostly true") than students in the Virtual treatment (Median = 4, "Somewhat true") on average ($Z = 8.86$, $p < 0.0001$; Fig. 2).

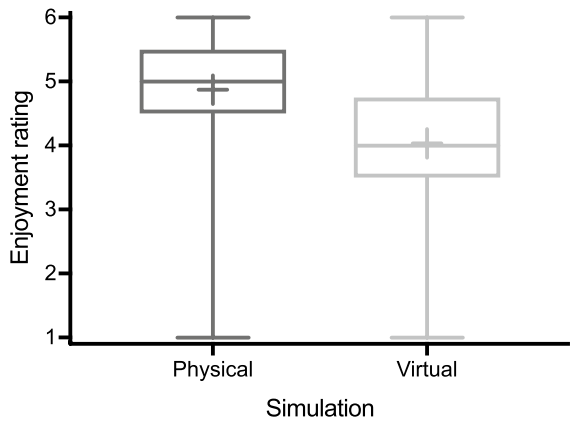

Fig. 2 Student self-reported enjoyment of laboratory activities. Students completed a survey after completing the lab activities and post-tests. They rated four statements about enjoying the lab on a scale of 1 (completely false) to 6 (completely true); their ratings across the four statements were averaged to calculate their overall enjoyment rating. *Centerlines* show the medians; *box limits* indicate the 25th and 75th percentiles; whiskers extend to the minimum and maximum values. Means are indicated by + symbol. Higher scores indicate higher reported levels of enjoyment. Students in the physical treatment reported significantly higher enjoyment

Student involvement in the lab activity differed by treatment

In addition to the students' self-report of how much they enjoyed the lab, we conducted classroom observations to look for observable correlates of student engagement, recording each student's behavior in four categories (Table 1). In each category, we were looking for evidence that students were involved in the lab, either through direct participation in the activity (as evidenced by verbal and motor behavior) or through active observation of the activity (as evidenced by direction of gaze and affect).

For each of the four behavior categories (gaze, verbal, motor, and affect), we compared the distributions of student behaviors into the relevant sub-categories between treatments using Chi squared tests (Fig. 3). There were significant differences between treatments in observed student behavior for all four behavior categories (gaze: $\chi^2 = 289.7$, $p < 0.0001$; verbal: $\chi^2 = 15.9$, $p < 0.01$; motor: $\chi^2 = 44.2$, $p < 0.0001$; affect: $\chi^2 = 96.1$, $p < 0.0001$). The Cramér's V estimate of the strength of association suggests a relatively strong association between treatment and gaze behavior ($V = 0.50$), a moderate association between treatment and affect ($V = 0.29$), and a weak

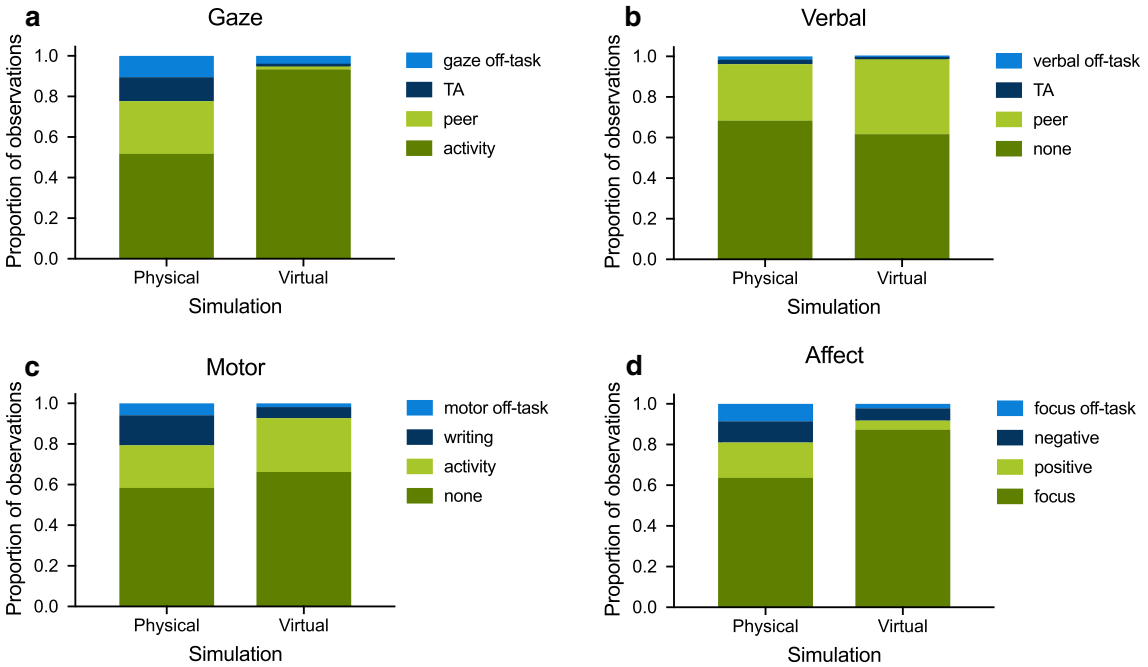

Fig. 3 Observations of student behavior while doing the physical simulation and virtual simulation lab activities. We conducted observations of student behavior in five lab sections for each treatment, over the course of 4 days (for details see "Methods"). Student gaze (**a**), verbal behavior (**b**), motor behavior (**c**), and affect (**d**) were recorded in mutually exclusive categories. There are significant differences between treatments in the distributions among behaviors in each category

association between treatment and verbal (V = 0.12) and motor behavior (V = 0.20; see Kotrlik and Williams 2003 for interpretations of Cramér's V).

The distributions between behaviors are most different between the two treatments in the gaze and affect categories (Fig. 3a, d). Notably, we saw more off-task behavior (usually texting or other cellphone activity) in students in the physical treatment, as can be seen in the gaze and motor behavior (Fig. 3a, c). Students in the virtual treatment overwhelming looked only at the computer screen, in contrast to the diversity of gaze directions in the physical treatment (Fig. 3a). The most common affect in both treatments was "focused" (appearance of interest in and focus on the lab activity), but this affect was more frequent in the virtual treatment. Students in the physical treatment were more likely to show "positive" affect (smiling or laughing), but were also more likely to show "negative" affect (frowning or appearance of boredom or frustration) and "focus off-task" (appearance of interest in and focus on something other than the lab activity, such as their cellphone, social conversation with peers or TA, or working on an overdue lab report).

There are less striking differences in behavioral distributions in the verbal and motor categories. Students in the virtual treatment talked more to their peers (Fig. 3b), although most students in both treatments showed no verbal behavior during observations. Surprisingly, we observed more activity-related motor behavior in the virtual treatments (typing on the keyboard, moving the mouse, or pointing at the screen; Fig. 3c), although students in the physical treatment spent more time writing (writing answers to the follow-up questions to be turned in individually after lab). This could be due to the fact that the group sizes in the labs doing the physical activity were larger, so although a few students in each group actively participated in the activities, many did not (although they may have participated in discussion).

The overall pattern suggests that most students were engaged by both the physical and the virtual lab activities, with a somewhat higher level of student involvement in the virtual treatment, given the higher proportion of students focused on the task and the lower proportion of off-task gaze, verbal, and motor behaviors.

Discussion

In this study we assessed student learning gains, self-reported enjoyment, and observed involvement with two natural selection simulations, one physical and the other virtual, as implemented in sections of a large-enrollment introductory biology laboratory course. The two simulations both attempt to address basic student misconceptions, and have the students act as active agents in the process of natural selection. Our results demonstrate that both simulations were effective in increasing students' knowledge of the key components of natural selection: variation, heritability, and differential reproductive success, based on their improved performance on a post-test and delayed post-test compared to their pre-test (Fig. 1). Student performance improved in both treatments, with no difference in student learning between the two treatments. However, we did find differences in students' self-reported enjoyment and their observed involvement with the activities. Overall, our results suggest that there is no clear-cut advantage to either the physical or virtual simulations we tested.

Student performance

In our study, students in both the physical and virtual treatments improved from the pre- to post-test from a mean of 71% to a mean of 84%. Both groups' scores declined a similar amount on the delayed post-test, but show evidence of retention of the concepts learned (Fig. 1). We also assessed student answers to short answer questions by scoring for the presence of particular concepts or misconceptions in each answer and comparing the frequency of concept and misconception use in the two treatments. Students in the two groups showed no significant differences in presence of concepts or misconceptions in their responses to the short answer questions (Tables 4, 5). Together these results suggest that both simulations are equally effective at improving student understanding, and help students grapple with the key concepts of evolution by natural selection.

The majority of students in both treatments included both key concepts and misconceptions in their responses to short answer questions, suggesting a "mixed mental model" of natural selection, which is common in introductory biology students even post-instruction (Nehm and Reilly 2007). Misconceptions about natural selection are persistent; we are more likely to overcome these with repeated instruction using a variety of approaches (Kalinowski et al. 2013). Our results suggest that either of these simulations can be a valuable tool for improving student understanding of natural selection in introductory courses, in combination with other types of instruction.

Self-reported enjoyment and observations of student involvement

Despite the similarity in outcomes, students differed in their reported enjoyment of the activities. To assess student enjoyment, we asked students to answer several questions using a Likert scale. Students in the physical treatment reported enjoying the activity significantly more than students in the virtual treatment (Fig. 2). The enjoyment rating of students in the virtual treatment

suggested that they also tended to enjoy that activity, but did not rate it as highly.

Of course, self-reported enjoyment gives only part of the picture. Whether students were engaged by the activity and what kinds of interactions occurred during the simulation also affects the success of an activity. We observed 10 lab sections (5 for each treatment) to assess student behavior—specifically gaze, affect, motor and verbal interactions.

Broadly, observations of both treatments showed most students were engaged in the simulation. However, students in the physical activity were involved in more diverse activities during the observations. During the physical simulation students moved about the room, and also had free time to socialize and engage in other off-task behaviors. These opportunities for participating in more varied activities, moving around more, and socializing may contribute to the higher enjoyment ratings of the physical simulations (Fig. 2).

Our quantification of student gaze dramatically demonstrates the difference in how students participated in the two lab activities. In over 90% of our observations of students in the virtual simulation, the students were looking directly at the screen. Students in the physical treatment were looking many different places: at peers (25%), at the TA (13%), or to a lesser extent off-task (<10%; Fig. 3a). This is not surprising since the physical activity involved more types of actions and only a portion of the students directly participated in the actual simulation (using the binder clips to "prey" on beans and beads) at any given time, while the rest participated as recorders or in discussions. The almost exclusive focus of the students' gaze in the virtual activity on the screen may have led to less viewing of their phones or other off-task items, which could be a unintended benefit of the activity. On the other hand, the variety of gaze direction in the physical activity could have other benefits. Chien et al. (2015) report a similar finding in a study where they recorded eye movements and gaze fixation of students using a lab activity based on a virtual simulation and a lab activity using physical materials (along with computer-aided data collection). They found that students participating in the virtual simulation activity showed longer fixation duration, which they suggested implies deeper cognitive processing. They also suggested that students in their virtual treatment could concentrate more on the relevant aspects of the task. Although our observation methods were much more coarse-grained and do not allow us to draw conclusions about what students' gaze was specifically focused on, we noted similar patterns.

Our observations of verbal and motor behavior show that most students in both treatments were not interacting verbally and not actively manipulating components of the activity during any given observation (Fig. 3b, c). Different patterns were seen in the less common verbal and motor behaviors. Students in the virtual activity were more likely to be speaking with their classmates. Not surprisingly, students in the virtual activity were more likely to be engaged in motor behavior relating to the simulation (manipulating the computer), whereas those in the physical activity were more likely to be writing (Fig. 3c), working on their answers to the post-lab questions. Interestingly, in their study, Chien et al. (2015) also found that students in the physical condition spent more time on their paper worksheet, even though the worksheet was identical between treatments; we find the same pattern here.

In our observations, the affect shown by a majority of students in both simulations was one of focused concentration (Fig. 3d). While the proportion of students showing focus was quite a bit higher in the virtual simulation, students in the physical activity were more likely to show a positive affect, consistent with the higher enjoyment ratings that the students gave to the physical activity (Fig. 2). Taken together, the positive and focused affects are less for the physical activity than the virtual activity, because students in the physical activity were more likely to be focused off-task or to show a negative affect (9 and 10% respectively). This may be a consequence of the fact that a greater proportion of the students in the physical simulation were not directly involved in the activity at any given time.

It is interesting that despite the greater level of off-task and negative affect, the enjoyment rating of the Physical activity was significantly higher. This points to a limitation in the self-reported enjoyment rating—we cannot distinguish between reported enjoyment being due to the activity itself or to the social setting of the activity, which allowed for more off-task behavior. However, over 80% of the students in the physical treatment exhibited focused or positive affect; combined with the demonstrated learning gains, this suggests that most students were both engaged in and learned from the activity.

Another limitation of both self-reported enjoyment and observations of student involvement is that they are likely to be highly context-dependent and therefore not generalizable to other classroom situations. However, they can provide insight into some best practices for implementing these activities. Some small changes in the implementation of both simulations could increase the proportion of students actively involved in either movement or discussion. Instructors in the physical activity could pose questions for groups to discuss after each round of the simulation, and both activities could have benefited from more whole-class discussion. The large group sizes in the physical sections limited involvement,

so where feasible, smaller groups would allow more students to actively participate at any given time; alternatively, instructors could ask students to switch roles so that more students have the opportunity to act as predator or recorder. Groups for the virtual activity could also be smaller, improving student motor involvement, but even in groups of 3–4, instructors can have students rotate control of the keyboard and mouse throughout the class period.

Similarities and differences in the activities that could impact learning gains

Student learning gains may be so similar between treatments because, although the simulations differ in instructional mode, the activities share similar learning objectives. They both actively illustrated that trait variation can lead to differential reproductive success, and that this can lead to change in phenotype frequency over time in a population. The student assumed the role of predator in both activities, although the participatory aspect of the simulation is much more limited in the virtual simulation. Both simulations also allowed students to work in groups, although group size differed substantially between the two activities. These broad similarities may account for the lack of difference observed in learning outcomes.

Despite these similarities, the activities do differ in many details, and given these differences, we expected we might see some differences in outcome in specific areas. One difference was that the virtual simulation includes many imbedded questions, with automatic feedback targeting misconceptions expressed in incorrect responses. Some questions involved visualization of data, whereas others asked students to reflect on what results meant. In class observations, we saw that these questions prompted student discussion within their groups, and these conversations were more common than in the physical treatment (Fig. 3b).

Because we used two simulations that are already in wide use, we did not carefully match the physical and virtual activities in this study. Thus, there may be different inherent advantages to each activity that nonetheless led to roughly equivalent learning through different routes. For example, benefits of the immediate feedback to questions in the virtual simulation may have been comparable to benefits from the movement-based and game-like participatory nature of the physical simulation. It is also possible that there may have been differences in learning that were not detectable with the assessments we used.

Differences in implementation of the activities that could affect student experience

Some important differences in the implementation of the two natural selection simulations in our study could have led to different student experiences, which could potentially impact both enjoyment and learning gains. The virtual activity included carefully worded step-by-step instructions, and feedback that encouraged students to repeat questions or activities until they have completed them correctly. The physical simulation included directions at the beginning, but students may have been able to alter portions of the procedure, unless a TA redirected them.

Based on our classroom observations, it appeared that the role of the teaching assistant tended to be quite different in the two activities. In the virtual activity, after helping the students access the module, most TAs then only answered occasional questions. In the physical activity, the TA was more actively involved in all stages of the activity. In addition, some of the TAs had taught the same lab course in previous years, and therefore had experience with the physical activity, whereas none had taught using the virtual simulation before. The TAs may have had different levels of enthusiasm for their activity; both the prior experience and the enthusiasm of the TA would likely influence the student experience of the activity. In our observations, the TAs for the physical activity were more likely to engage the whole section in general discussion; TAs for the virtual simulation may have benefited from more direction from course instructors on how to incorporate more whole-class discussion in the activity.

While we did survey students about their enjoyment of the two activities, we did not ask them to self-assess their own learning gains or their opinions about the value of the activity. Had we done so, we may have uncovered other student opinions about the lab in addition to simply enjoyment (Pyatt and Sims 2007; Wiesner and Lan 2004). In a study comparing physical engineering lab activities to remote and simulated labs (both conducted by students outside of the lab classroom), Corter et al. (2007) found that students varied in their expressed preferences for different types of lab, and their preferences were related to whether they valued factors like in-person group interactions or the ease of data acquisition; they did not find any significant correlations of student preferences with predictors such as GPA or SAT scores. Anecdotally, we heard some students and TAs express that they felt the physical activity was too juvenile and not appropriate for a college lab; however, many students clearly enjoyed themselves, and opinions about the value of physical manipulation and the game-like competition of the physical simulation are likely to vary substantially among students.

Choosing a natural selection simulation activity

Our results align with other recent studies finding no difference in learning gains between physical and virtual

activities (Chen et al. 2014; Corter et al. 2007; Hawkins and Phelps 2013; Kelly et al. 2008; Klahr et al. 2007; Lazonder and Ehrenhard 2014; Tatli and Ayas 2013), suggesting there is no evidence that the physical or virtual nature of an activity in itself confers a benefit in terms of learning gains. Thus future research should start to focus more on other aspects of student experience so that we can learn more about the contexts in which one or the other type of activity might be preferable.

Given the equivalent learning gains we found in the physical and virtual natural selection simulations we studied, decisions about which activity to use can focus on what is most useful and feasible in a specific classroom context. Each of the simulations we studied has benefits and limitations. Students reported enjoying the physical activity more, but they were also more likely to engage in off-task behaviors, most likely due to the necessarily larger group size—this in itself might make the activity worth avoiding, depending on the specific student population and classroom environment. This finding points to what may be a shortcoming in participatory simulations in general—benefits may be more likely to accrue to the participants than to the bystanders, so careful design and implementation of activities can maximize the proportion of students who participate. On the other hand, the physical activity may be helpful to some students who will enjoy or learn better from the social and tactile interactions in such a task (Corter et al. 2007). We feel the activity could be improved by more frequent breaks for reflection, answering questions, and discussion as a class.

The virtual activity has the distinct advantage of not requiring a large group, space, or materials, although it clearly requires access to computers and is susceptible to technology failures. It could be performed as an out-of-class activity, although this could then reduce the known benefits of group discussion (Linton et al. 2014; Smith et al. 2009; Stamovlasis et al. 2006), unless implemented in a way to allow out-of-class asynchronous group work (e.g., Corter et al. 2007). Virtual simulations in general can be more flexible and allow for students to view events from multiple perspectives (National Research Council 2011; Rutten et al. 2012; Smetana and Bell 2012). Some previous authors have pointed to simulations being less costly than some comparable physical lab activities that require expensive equipment, supplies, or animals (Dewhurst et al. 1994; Wiesner and Lan 2004), but this was not a concern with the physical natural selection simulation used here as it uses cheap and readily available materials. Virtual simulations in some cases can be a more efficient use of class time (Gibbons et al. 2004; Pyatt and Sims 2007; Trundle and Bell 2010).The implementation of the virtual activity in a lab setting could be made more effective with more support and direction for TAs about how they might more actively guide the lab and interact with the students, such as suggesting points in the simulation where they might pause for a class-wide discussion, and providing discussion questions and ideas for whole-class data collection and comparison of results. These changes would likely increase the satisfaction of both students and TAs with the activity.

Conclusions

When choosing which active learning strategy to use for natural selection, instructors must necessarily base part of their decision making on available resources and financial considerations; but of course, the relative effectiveness of a given activity for achieving the desired learning objectives should be a key consideration (Freeman et al. 2014). In our study, we assessed two popular simulations of natural selection and found them both to be effective, and saw no differences in learning gains between the simulations. Both of these activities are suitable choices for improving student understanding of natural selection. Depending on the particular circumstances and learning objectives, one or the other might prove more effective. The results from our student survey and classroom observations may help inform considerations about how the simulation might work in a specific context.

Although in our study the two simulations we compared show similar learning gains, it is unlikely that all activities or strategies for teaching natural selection will prove equally helpful (Nehm 2006). Furthermore, as argued by Freeman et al. (2014), more comparisons of alternative active learning approaches will lead to an improved understanding of the specific contexts and populations for which different activities are likely to be most effective.

Additional file

Additional file 1. Additional material.

Abbreviations

TA: teaching assistant.

Authors' contributions

CMR originally conceived of the study; all three authors planned the research. CMR coordinated all of the TAs and labs, implemented and collected data from all online tests and surveys; DSP and JCM observed lab classes; CMR and DSP coded student post-lab responses to short answer questions; DSP and JCM conducted data analysis. All three authors wrote and edited the manuscript. All authors read and approved the final manuscript.

Author details

[1] SimBio, 1280 S 3rd St W, Suite 3, Missoula, MT 59801, USA. [2] Department of Biology, University of Massachusetts, Amherst, MA 01003, USA. [3] Department of Instructional Technology and Learning Sciences, University of Utah, Logan, UT 84321, USA. [4] Present Address: Center for the Integration of Research, Teaching and Learning, 1025 W. Johnson St., Madison, WI 53706, USA.

Acknowledgements
We thank all of the students and teaching assistants in the course for their cooperation with this study and for allowing us to observe their classrooms, and for the lab managers, Robert Cairl and David Prodana for their help with organization and implementation. Susan Maruca, Joel Abraham, W. John Roach, Eli Meir and anonymous reviewers provided comments that improved the manuscript.

Competing interests
DSP and JCM were involved in the development of the revised version of the Darwinian Snails software module. DSP was an employee of SimBio when this study was conducted, and DSP and JCM are collaborating with SimBio on several projects.

Ethical approval and consent to participate
This study was approved by the University of Massachusetts Amherst Institutional Review Board (protocol # 2013-1885). Students consented to participate through an online form that preceded the first assessment. For our analysis of the assessments and survey, we only used data from students who consented; for the classroom observations, observers did not know student identity and the IRB waived informed consent.

Funding
This research was partially supported by NSF (IIS-1227245).

Publisher's Note

Received: 7 February 2017 Accepted: 6 July 2017
Published online: 14 July 2017

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. doi:10.1007/s12052-009-0142-3.

Alters BJ, Nelson CE. Perspective: teaching evolution in higher education. Evolution. 2002;56(10):1891–901.

Altmann J. Observational study of animal behavior: sampling methods. Behaviour. 1974;49(3/4):227–67.

American Association for the Advancement of Science. AAAS science assessment website. Washington: American Association for the Advancement of Science; 2013. http://assessment.aaas.org/.

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. doi:10.1002/tea.10053.

Bishop BA, Anderson CW. Student conceptions of natural selection and its role in evolution. J Res Sci Teach. 1990;27(5):415–27. doi:10.1002/tea.3660270503.

BraySpeth E, Long TM, Pennock RT, Ebert-May D. Using Avida-ED for teaching and learning about evolution in undergraduate introductory biology courses. Evol: Educ Outreach. 2009;2(3):415–28. doi:10.1007/s12052-009-0154-z.

Cheetham AH, Hazel JE. Binary (presence-absence) similarity coefficients. J Paleontol. 1969:1130–36.

Chen S, Chang W-H, Lai C-H, Tsai C-Y. A comparison of students' approaches to inquiry, conceptual learning, and attitudes in simulation-based and microcomputer-based laboratories. Sci Educ. 2014;98(5):905–35. doi:10.1002/sce.21126.

Chien KP, Tsai CY, Chen HL, Chang WH, Chen S. Learning differences and eye fixation patterns in virtual and physical science laboratories. Comput Educ. 2015;82:191–201. doi:10.1016/j.compedu.2014.11.023.

Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Erlbaum Associates; 1988.

Corter JE, Nickerson JV, Esche SK, Chassapis C, Im S, Ma J. Constructing reality: a study of remote, hands-on, and simulated laboratories. ACM Trans Comput-Hum Interact. 2007;14(2):7. doi:10.1145/1275511.1275513.

Cramér H. Mathematical methods of statistics. Princeton: Princeton University Press; 1946.

Deci EL, Eghrari H, Patrick BC, Leone DR. Facilitating internalization: the self-determination theory perspective. J Pers. 1994;62:119–42.

Dewhurst DG, Hardcastle J, Hardcastle PT, Stuart E. Comparison of a computer simulation program and a traditional laboratory practical class for teaching the principles of intestinal absorption. Am J Physiol. 1994;267:S95.

Eddy SL, Converse M, Wenderoth MP. PORTAAL: a classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. CBE-Life Sci Educ. 2015;14:ar23. doi:10.1187/cbe.14-06-0095.

Eterovic A, Santos CMD. Teaching the role of mutation in evolution by means of a board game. Evol: Educ Outreach. 2013;6:22. doi:10.1186/1936-6434-6-22.

Fifield S, Fall B. A hands-on simulation of natural selection in an imaginary organism, Platysoma apoda. Am Biol Teach. 1992;54(4):230–5.

Finkelstein ND, Adams WK, Keller CJ, Kohl PB, Perkins KK, Podolefsky NS, LeMaster R. When learning about the real world is better done virtually: a study of substituting computer simulations for laboratory equipment. Phys Rev Spec Top-Phys Educ Res. 2005;1(1):010103. doi:10.1103/PhysRevSTPER.1.010103.

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. doi:10.1073/pnas.1319030111.

Gibbons NJ, Evans C, Payne A, Shah K, Griffin DK. Computer simulations improve university instructional laboratories. CBE-Life Sci Educ. 2004;3(4):263–9. doi:10.1187/cbe.04-06-0040.

Gregory TR. Understanding natural selection: essential concepts and common misconceptions. Evol: Educ Outreach. 2009;2(2):156–75. doi: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.

Hawkins I, Phelps AJ. Virtual laboratory vs traditional laboratory which is more effective for teaching electrochemistry. Chem Educ Res Pract. 2013;14(4):516–23. doi:10.1039/C3RP00070B.

Herron JC, Maruca S, Meir E. Darwinian snails. Missoula: SimBio; 2014. http://www.simbio.com.

Hildebrand TJ, Govedich FR, Bain BA. Hands-on laboratory simulation of evolution: an investigation of mutation, natural selection, & speciation. Am Biol Teach. 2014;76(2):132–6.

Janulaw A, Scotchmoor J. Clipbirds. Understanding evolution. Berkeley: University of California Museum of Paleontology; 2011. http://www.ucmp.berkeley.edu/education/lessons/clipbirds/. Accessed 22 Sept 2014.

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. doi:10.1187/cbe-12-06-0070.

Kelly J, Bradley C, Gratch J. Science simulations: do they make a difference in student achievement and attitude in the physics laboratory? (Evaluative Report No. ED501653). Educational Resources Information Center (ERIC); 2008. http://eric.ed.gov/, (http://eric.ed.gov/?id=ED501653).

Klahr D, Triona LM, Williams C. Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. J Res Sci Teach. 2007;44(1):183–203. doi:10.1002/tea.20152.

Kotrlik JW, Williams HA. The incorporation of effect size in information technology, learning, and performance research. Inf Technol Learn Perform J. 2003;21(1):1–7.

Krist AC, Showsh SA. Experimental evolution of antibiotic resistance in bacteria. Am Biol Teach. 2007;69(2):94–7. doi:10.1662/0002-7685(2007).

Latham LG, Scully EP. Critters! A realistic simulation for teaching evolutionary biology. Am Biol Teach. 2008;70(1):30–3.

Lazonder AW, Ehrenhard S. Relative effectiveness of physical and virtual manipulatives for conceptual change in science: how falling objects fall. J Comput Assist Learn. 2014;30(2):110–20. doi:10.1111/jcal.12024.

Linton DL, Farmer JK, Peterson E. Is peer interaction necessary for optimal active learning? CBE-Life Sci Educ. 2014;13(2):243–52. doi:10.1187/cbe.13-10-0201.

Marshall P, Cheng PCH, Luckin R. Tangibles in the balance: a discovery learning task with physical or graphical materials. In Proceedings of the fourth

international conference on Tangible, embedded, and embodied interaction. New York: Association for Computing Machinery; 2010. p. 153–160. http://dl.acm.org/citation.cfm?id=1709914.
Marx JD, Cummings K. Normalized change. Am J Phys. 2007;75(1):87–91.
National Research Council. Learning science through computer games and simulations. In: Honey MA, Hilton M, editors. Committee on science learning: computer games, simulations, and education. Washington: Board on Science Education, Division of Behavioral and Social Sciences and Education, The National Academies Press. http://www.nap.edu/catalog/13078.
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.
Nelson CE. Teaching evolution (and all of biology) more effectively: strategies for engagement, critical reasoning, and confronting misconceptions. Integr Comp Biol. 2008;48(2):213–25. doi:10.1093/icb/icn027.
Nickerson JV, Corter JE, Esche SK, Chassapis C. A model for evaluating the effectiveness of remote engineering laboratories and simulations in education. Comput Educ. 2007;49(3):708–25. doi:10.1016/j.compedu.2005.11.019.
Plunkett AD, Yampolsky LY. When a fly has to fly to reproduce: selection against conditional recessive lethals in Drosophila. Am Biol Teach. 2010;72(1):12–5. doi:10.1525/abt.2010.72.1.4.
Price RM. Performing evolution: role-play simulations. Evol: Educ Outreach. 2010;4:300. doi:10.1007/s12052-010-0300-7.
Pyatt K, Sims R. Learner performance and attitudes in traditional versus simulated laboratory experiences. Proceedings ICT: providing choices for learners and learning ascilite. Singapore: Australasian Society for Computers in Learning in Tertiary Education; 2007. http://www.academia.edu/download/30655038/pyatt.pdf.
Real R, Vargas JM. The probabilistic basis of Jaccard's index of similarity. Syst Biol. 1996;45(3):380. doi:10.2307/2413572.
Renken MD, Nunez N. Computer simulations and clear observations do not guarantee conceptual understanding. Learn Instr. 2013;23:10–23. doi:10.1016/j.learninstruc.2012.08.006.
Rutten N, van Joolingen WR, van der Veen JT. The learning effects of computer simulations in science education. Comput Educ. 2012;58(1):136–53. doi:10.1016/j.compedu.2011.07.017.
Sawada D, Piburn MD, Judson E, Turley J, Falconer K, Benford R, Bloom I. Measuring reform practices in science and mathematics classrooms: the reformed teaching observation protocol. School Sci Math. 2002;102(6):245–53.
Scoville SA, Buskirk TD. Traditional and virtual microscopy compared experimentally in a classroom setting. Clin Anat. 2007;20(5):565–70. doi:10.1002/ca.20440.
Seeley RH. Intense natural selection caused a rapid morphological transition in a living marine snail. Proc Natl Acad Sci. 1986;83(18):6897–901.
Serafini A, Matthews DM. Microbial resistance to triclosan: a case study in natural selection. Am Biol Teach. 2009;71(9):536–40. doi:10.1662/005.071.0907.
Siegel MA, Mlynarczyk-Evans S, Brenner TJ, Nielsen KM. A natural selection: partnering teachers and scientists in the laboratory creates a dynamic learning community. Sci Teach. 2005;72(7):42–5.
Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences (Second). Boston: McGraw-Hill; 1988.
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. doi:10.1080/09500693.2011.605182.
Smith MK, Jones FHM, Gilbert SL, Wieman CE. The classroom observation protocol for undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom practices. CBE-Life Sci Educ. 2013;12(4):618–27. doi:10.1187/cbe.13-08-0154.
Smith MK, Wood WB, Adams WK, Wieman C, Knight JK, Guild N, Su TT. Why peer discussion improves student performance on in-class concept questions. Science. 2009;323(5910):122–4. doi:10.1126/science.1165919.
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. doi:10.1080/09500690110095285.
Stamovlasis D, Dimos A, Tsaparlis G. A study of group interaction processes in learning lower secondary physics. J Res Sci Teach. 2006;43(6):556–76. doi:10.1002/tea.20134.
Steinberg RN. Computers in teaching science: to simulate or not to simulate? Am J Phys. 2000;68(S1):S37–41.
Taghavi SE, Colen C. Computer simulation laboratory instruction vs. traditional laboratory instruction in digital electronics. J Inf Technol Impact. 2009;9(1):25–36.
Tatli Z, Ayas A. Effect of a virtual chemistry laboratory on students' achievement. Educ Technol Soc. 2013;16(1):159–70.
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. doi:10.1187/cbe-13-07-0136.
Trundle KC, Bell RL. The use of a computer simulation to promote conceptual change: a quasi-experimental study. Comput Educ. 2010;54(4):1078–88. doi:10.1016/j.compedu.2009.10.012.
Van Thiel LR. Predator-prey coevolution. In Goldman CA, editors. Tested studies for laboratory teaching. Proceedings of the 15th Workshop/Conference of the Association for Biology Laboratory Education (ABLE). vol. 15; 1994. p. 293–318.
Wiesner TF, Lan W. Comparison of student learning in physical and simulated unit operations experiments. J Eng Educ. 2004;93(3):195–204.
Yamanoi T, Iwasaki WM. Origami bird simulator: a teaching resource linking natural selection and speciation. Evol: Educ Outreach. 2015. doi:10.1186/s12052-015-0043-6.