A CURE on the Evolution of Antibiotic Resistance in *Escherichia coli* Improves Student Conceptual Understanding

Scott Freeman,[†*] Joya Mukerji,[†‡] Matt Sievers,[†] Ismael Barreras Beltran, Katie Dickinson, Grace E. C. Dy, Amanda Gardiner, Elizabeth H. Glenski, Mariah J. Hill, Ben Kerr, Deja Monet, Connor Reemts, Elli Theobald, Elisa T. Tran, Vicente Velasco, Lexi Wachtell, and Liz Warfield
Department of Biology, University of Washington, Seattle, WA 98195

ABSTRACT

We developed labs on the evolution of antibiotic resistance to assess the costs and benefits of replacing traditional laboratory exercises in an introductory biology course for majors with a course-based undergraduate research experience (CURE). To assess whether participating in the CURE imposed a cost in terms of exam performance, we implemented a quasi-experiment in which four lab sections in the same term of the same course did the CURE labs, while all other students did traditional labs. To assess whether participating in the CURE impacted other aspects of student learning, we implemented a second quasi-experiment in which all students either did traditional labs over a two-quarter sequence or did CURE labs over a two-quarter sequence. Data from the first experiment showed minimal impact on CURE students' exam scores, while data from the second experiment showed that CURE students demonstrated a better understanding of the culture of scientific research and a more expert-like understanding of evolution by natural selection. We did not find disproportionate costs or benefits for CURE students from groups that are minoritized in science, technology, engineering, and mathematics.

INTRODUCTION

The National Science Foundation and Howard Hughes Medical Institute have promoted course-based undergraduate research experiences (CUREs) as a tool for increasing the representation of female, first-in-family with a 4-year degree (1st-gen), underrepresented minority (URM), and low-socioeconomic status (SES) students in science, technology, engineering, and mathematics (STEM; Wei and Woodin, 2011; Elgin *et al.*, 2016; Estrada *et al.*, 2016). This effort is motivated by two observations: The literature documents better retention of STEM-interested students who participate in classical, apprentice-style undergraduate research experiences (UREs; Gentile *et al.*, 2017; Wilson *et al.*, 2018), but students typically do UREs in their junior and senior years, after many STEM-interested but minoritized students have changed majors or dropped out of college—often due to poor performance or intellectually and emotionally unrewarding experiences in large-enrollment introductory courses (Herrera and Hurtado, 2011; National Academies of Sciences, Engineering, and Medicine, 2016; Harris *et al.*, 2020). In addition, UREs that lack a stipend may act as a barrier to participation for low-income students, and unpaid internships can have negative impacts on low-SES individuals who do participate (McHugh, 2016). CUREs are being endorsed as a tool to resolve these conflicts and democratize access to undergraduate research (Bangera and Brownell, 2014; Elgin *et al.*, 2016).

Early work on CUREs focused on clarifying which aspects of a lab's design qualify it as an authentic research experience. An emerging consensus focused on four key attributes: 1) use of scientific practices; 2) collaboration; 3) iteration, defined as

Erin L. Dolan, *Monitoring Editor*
Submitted Dec 7, 2021; Revised Nov 8, 2022; Accepted Dec 2, 2022
CBE Life Sci Educ March 1, 2023 22:ar7
DOI:10.1187/cbe.21-12-0331

[†]Co–first authors, in alphabetical order.
[‡]Present address: Department of Biological Sciences, California State University–Sacramento, Sacramento CA 95819.

Conflict of interest statement: S.F., J.M., K.D., B.K., and L.W. helped develop the CURE curriculum. The authors warrant that no promotion of a particular product, to the exclusion of other similar products, should be construed.

*Address correspondence to: Scott Freeman (srf991@uw.edu).

building on or replicating previous work; and 4) making discoveries that are relevant to the scientific community (Auchincloss *et al.*, 2014; Brownell and Kloser, 2015; Cooper *et al.*, 2019). It is important to note that labs in which student teams design experiments and collect data can fulfill the first three criteria and lead to improved student outcomes (Luckie *et al.*, 2004; Rodrigo-Pieris *et al.*, 2018; Indorf *et al.*, 2019). Although labs like these qualify as inquiry experiences, they are not necessarily CUREs, because they lack the fourth attribute—producing data that are relevant to the scientific community or even potentially publishable (Wiley and Stover, 2014; Cooper *et al.*, 2017; Hanauer *et al.*, 2017). This is a critical distinction, as producing data that are both novel and of interest to the scientific community can improve students' sense of cognitive and emotional ownership in the experience (Cooper *et al.*, 2019). Both aspects of ownership are associated with increased student intention to pursue a research career (Corwin *et al.*, 2018).

Some CUREs have been integrated into courses as the lab component, while others are offered as stand-alone courses. To date, many published CUREs have been focused on discovery science, such as screening mutants (Gasper and Gardner, 2013), characterizing biodiversity via barcoding or environmental sequencing (Jacob, 2012; Harris and Bellino, 2013; Wang *et al.*, 2015), isolating and describing uncharacterized types of phage (Hatfull, 2010; Jordan *et al.*, 2014), documenting allelic variation at a specific gene (Lau and Robinson, 2009), or annotating genomes (Shaffer *et al.*, 2010; Burnette and Wessler, 2013). While acknowledging the value of discovery science, researchers have also been searching for ways to design CUREs with a stronger experimental emphasis (Hatfull, 2010).

Whether they are focused on discovery or experimental science, most CUREs are motivated by the hypothesis that the work will encourage students to persist in STEM. As a result, researchers have focused on three general categories of outcome variables when assessing a CURE's impact: 1) measures of learning, 2) attitudes that are known to support STEM retention, and 3) persistence to graduation in STEM (Rodenbusch *et al.*, 2016; Corwin *et al.*, 2018; Cooper *et al.*, 2019; Indorf *et al.*, 2019). Here, we focused on understanding changes in aspects of student learning.

Because CUREs have existed for fewer than 15 years, research on their impact is "still in the early stages of development" (Gentile *et al.*, 2017, pS-4). Along with other workers (Indorf *et al.*, 2019; Krim *et al.*, 2019), we note that some early studies lacked a comparison group, lacked controls over student characteristics when a comparison group was used, used self-reported learning gains instead of objective assessments, lacked pre data for measuring learning gains, lacked controls for motivation and self-selection in studies of "opt-in" CUREs, or analyzed aggregate instead of by-student data.

Researchers have used several approaches to solve the difficult task of assessing CUREs rigorously. Propensity-score matching and other regression-based approaches use data on demographics and past performance to control for potential differences between comparison groups (e.g., Rodenbusch *et al.*, 2016; Hanauer *et al.*, 2017, 2022; Indorf *et al.*, 2019). A growing body of work is producing instruments with at least some validity evidence for studying CURE instruction (Hanauer and Dolan, 2014; Corwin *et al.*, 2015), and recent longer-term studies have measured graduation rates in STEM fields and other particularly meaningful outcomes (e.g., Rodenbusch *et al.*, 2016). Pre–post testing regimes, with pre scores used as a fixed effect in regression models, have also become standard in the field. Although no study of an opt-in CURE has used a randomized controlled trial to control for self-selection bias—with student volunteers randomly assigned to a CURE or control treatment (e.g., Wischusen and Wischusen, 2007; Stanich *et al.*, 2018)—CUREs that are integrated into existing courses can be studied using quasi-random designs.

The goal of our work was to assess the costs and benefits of replacing traditional laboratory exercises in an introductory biology course for majors with a CURE. To build on the existing literature, the CURE design blended discovery and experimental science and emphasized scientific practices, collaboration, iteration, and producing potentially publishable data. In addition, the study employed quasi-experimental designs, pre–post testing, and data on student characteristics in assessing two outcomes:

1. *course performance* as measured by scores on identical exams; and
2. *other measures of learning*—specifically, student understanding of the culture of scientific research (CSR), evolution by natural selection, and experimental design.

We analyzed data on these measures for the overall student population as well as for four demographic groups that are minoritized in STEM: women, URMs, 1st-gen students, and low-SES students.

METHODS

Course Context

The study took place in the first two quarters of a three-quarter, yearlong introductory sequence for majors at the University of Washington (UW). Course 1 (Biology 180) introduces experimental design, evolution, Mendelian genetics, diversity of life, and ecology. Course 2 (Biology 200) introduces biological molecules, molecular genetics, cell biology, and developmental biology. (A third course, focused on animal and plant physiology, completes the yearlong sequence.) Total course enrollment ranged from 500 to 1200, depending on the term in question. The laboratory component of each course is required and meets once per week for a minimum of 1 hour and 50 minutes, with each section enrolling 24 students. Labs account for about 10% of overall course points and are taught by graduate teaching assistants (TAs) who are trained by a course coordinator.

Although each course in the sequence is taught in a high-structure format with intensive active learning during class sessions (Haak *et al.*, 2011), the labs are more traditional. They can be categorized as 1) workshop-style, often pencil-and-paper exercises that are designed to deepen student understanding of particularly difficult course concepts; or 2) inquiry experiences that allow students to ask a question, formulate a hypothesis, collect data, and analyze data to test predictions. Although one Course 1 lab produced data that resulted in a publication (Freeman *et al.*, 2016), that exercise was not designed to do so. During the period of this study, all of the traditional laboratory exercises had predetermined outcomes and had been taught many times, over many years.

The CURE Intervention

The CURE focused on experimental evolution of antibiotic resistance in *Escherichia coli*. The scientific question and protocols were developed by a faculty member (B.K.) and his lab group as an extension of their research program and were designed to address questions of current interest in the literature: specifically, the evolution of cross-drug effects and compensatory mutations in antibiotic-resistant strains. Cross-drug or collateral effects occur when strains that are resistant to one drug show increased sensitivity or resistance to one or more other drugs (Tekin *et al.*, 2018); compensatory mutations lower the fitness cost of resistance and help alleles for resistance persist in populations even in the absence of the drug. These issues are clinically relevant, because compensatory mutations, cross-resistance, and collateral sensitivity impact the design of drug cocktails, drug cycling, and other therapeutic regimes.

The CURE was designed to mimic a classical, apprentice-style URE as closely as possible. Specifically, students did not create the CURE's scientific question, hypothesis, experimental design, or protocols. Instead, students implemented established protocols with help from graduate TAs and peer facilitators, then collected and analyzed data that are potentially publishable. As a result, students' sense of ownership in the project depended on their intellectual and emotional connection to the question and to their data (Hanauer *et al.*, 2012; Hanauer and Dolan, 2014) as opposed to creating the question and experimental design (Olimpo *et al.*, 2016; Indorf *et al.*, 2019).

CURE labs replaced the first seven of nine lab sessions in Course 1. CURE sessions replaced five traditional labs in Course 2 but were scheduled intermittently, so the remaining traditional labs could be synchronized with lecture material.

Supplemental Table S1 summarizes the sequence of 12 CURE lab sessions, and a detailed description of the protocols is published elsewhere (Dickinson *et al.*, 2021). During both experiments reported here, Course 1 students worked in groups of four to expose a sample of *E. coli* to the antibiotic rifampicin and another sample to the antibiotic streptomycin. Because each group isolated its own strains that were resistant to each antibiotic, each group was working with unique strains—even though all groups performed the same procedures throughout. After this isolation step, each group performed daily transfers that allowed each of the resistant strains—as well as sensitive strains propagated as a control—to evolve in the absence of antibiotics. These transfers were done on a drop-in basis, outside each lab group's normally scheduled weekly lab session. In addition, the groups assayed the relative fitness and level of drug resistance of each ancestral and descendant strain from the beginning and end of the daily transfers, respectively. In Course 2, groups amplified and sequenced a candidate gene that is often mutated in resistant forms, visualized the predicted 3D structure of the protein product, and produced and presented a poster that synthesized the molecular data with the data on fitness and level of resistance. It is important to note that these protocols were designed to allow the experimental design to change over time in response to results generated by students, for example, by varying the antibiotics being studied, the conditions for the experimental evolution step (daily transfers), and the target of gene sequencing.

The curriculum emphasized the relevance of the data to important scientific questions. For example, increased fitness in descendant versus ancestral resistant strains is consistent with the existence of compensatory mutations. If streptomycin resistance changed in strains that were selected for rifampicin resistance, it would be evidence for cross-resistance or collateral sensitivity. In addition, students understood that their data were archived in a database that would be mined by advanced undergraduates or other Kerr lab members interested in the phenotypic consequences of specific mutations that impact antibiotic resistance.

Two Experimental Approaches for Studying Student Outcomes

We organized a concurrent experiment to test the hypothesis that participating in the CURE would lead to lower exam performance, due to a trade-off with traditional labs in terms of reinforcing student understanding of core concepts. In this design, four lab sections in Course 1 were designated as CURE sections. All four sections met at times that were average in terms of how rapidly they filled as students registered for the course—meaning that they were not high-demand or low-demand times. Students in the remaining 19 lab sections did the traditional exercises, but we only analyzed outcomes in the four sections that met at the same time on the same day as the CURE sections, as an additional control on variation in student preparation, ability, and degree of undergraduate experience, because registration slots are assigned based on accumulated credit hours. All other aspects of the courses were identical.

Although we carried the concurrent experiment through to Course 2 with three experimental sections doing the CURE labs, only 16 students out of the 96 who originally enrolled in Course 1 CURE sections took Course 2 in the subsequent term and had course schedules that allowed them to register for those CURE sections. Because most students in the CURE treatment only completed seven of the 12 CURE labs, the only data we report from the concurrent experiment are exam scores during Course 1.

The longitudinal study took place over an academic year. We did pre–post testing of students who 1) completed traditional labs in the Autumn and Winter terms for Course 1 and Course 2 or 2) did CURE labs in Winter and Spring terms for Course 1 and Course 2. This schedule was chosen purposefully. Anecdotally, course staff note that students who take the sequence "on-track," in Fall and Winter, are better prepared and more motivated, on average, than students who start the series later in the academic year. The Winter–Spring sequence also has a much higher percentage of first-year students than the Fall–Winter cohort, meaning that their counterparts in the Fall–Winter terms have more experience as college students. Thus, the experiment's schedule was designed to bias the outcome against the hypothesis that CUREs are beneficial.

In the longitudinal experiment, the comparison groups experienced different course instructors and took different exams. Although the longitudinal design was not as tightly controlled as the concurrent design, the sample size was large enough to allow us to investigate whether the CURE might have disproportionately large benefits for four underrepresented groups in STEM: women, low-SES, 1st-gen, and URM.

Data Collection

In addition to collecting data on course performance and student demographics, we asked students to fill out a survey at the

start and/or end of each course. The online instrument opened for 48 hours during the first week and the last week of the course, and students were given a small number of course points for completing it at each instance.

Whenever possible, we used open-response questions from published instruments that addressed constructs of interest in this study and that had validity evidence. But we also followed Harrison *et al.* (2011) and Irby *et al.* (2018) in developing assessments of interest for the CURE in question.

Independent Variables

The independent variables in the longitudinal study were treatment as CURE or non-CURE, pre score on the dependent variable in question if relevant, and college entrance examination total score as fixed effects. We also included status in one of the four underrepresented groups as both a fixed effect and as an interaction term with treatment. We coded student characteristics—gender, URM status, SES status, and family educational status—as binary variables based on data obtained from the UW registrar's office. We used total Scholastic Aptitude Test (SAT) score or total ACT score converted to the SAT scale using concordance tables published by the College Board to control for variation in academic preparation or ability. In addition to SAT score, some models included course grade as an additional control for the variation in student ability and preparation that exists among lab sections and terms.

Dependent Variables

Course Performance. During the concurrent experiment, we measured scores on identical exam questions to test the hypothesis that replacing existing labs with CURE labs would exact a cost in terms of exam performance, as several of the traditional labs were designed to deepen student understanding of concepts that regularly are tested on exams, while CURE-specific questions did not appear on exams. The Course 1 lab coordinator, who was blind to the rationale behind this analysis, identified exam questions that were or were not directly relevant to material covered in the traditional labs. The points that were directly relevant to the traditional labs totaled 53 on exam 1, 32 on exam 2, 28 on exam 3, and 13 on exam 4; the lab coordinator identified no exam questions that were specifically relevant to the CURE. We used by-student scores totaled for the non-lab questions as a predictor in a regression model and tested whether scores on questions addressed in the traditional labs differed among the two treatments in each of the four 100-point exams.

Other Measures of Learning. We did not evaluate exam scores during the longitudinal study, because test questions were not identical between treatments. The constructs assessed in the longitudinal experiment included three other measures of learning besides exam scores; however, each was assessed via open-response questions that were included in the pre- and postcourse surveys. The surveys are provided in the Supplemental Material.

1. **Culture of scientific research.** We used the prompts "What does it mean to think like a scientist?" (Brownell *et al.*, 2015), "What does it mean to do science?," and "Did you do real science in your [course name] labs?" to document student understanding of the culture of scientific research (Dewey *et al.*, 2021). We were interested in quantifying student progress in understanding what scientists do, value, and believe—in essence, what it means to be a scientist—because both guided inquiry and CURE labs are motivated in part by the goal of helping students develop scientific habits of mind and practice (Brownell *et al.*, 2015). The "real research" prompt appeared only on the post survey for each treatment, as it explicitly referred to students' lab experience in the focal course. The scoring rubrics for the three prompts were developed by Wachtell, L., Gardiner, A., Sievers, M., Dickinson, K., Dy, G.E.C., Glenski, E.H., Mukerji, J., Theobald, E., Tran, E.T., Velasco V., and Freeman, S. (unpublished data) and are provided in Supplemental Table S2.
2. **Experimental design.** We quantified student understanding of experimental design, because one of the CURE's fundamental design goals was to emphasize experimental over discovery science. To do so, we used the published expanded experimental design ability tool (E-EDAT) prompts and scoring rubric (Supplemental Table S3; see also Brownell *et al.*, 2014). As the surveys provided in Supplemental Material indicate, students were presented with different E-EDAT prompts on the pretest at the start of Course 1 and the posttest at the end of Course 2. We did this to minimize the possibility of measuring artificial changes in students' experimental design ability due to familiarity.
3. **Evolution by natural selection.** We evaluated student understanding of natural selection, because it is one of five core concepts in *Vision and Change* (American Association for the Advancement of Science, 2011), a major learning outcome for the introductory series we studied, and central to the scientific questions addressed in the CURE. We documented student learning using modified forms of the assessing contextual reasoning about natural selection (ACORNS) prompts, which ask students to explain how a specified novel trait evolved from a specified ancestral state (Nehm *et al.*, 2012), and an updated scoring rubric called the E-ACORNS (Sievers *et al.*, 2023). An example of an E-ACORNS prompt is: "One species of prosimians (animals) has long tarsi. How would biologists explain how this species with long tarsi evolved from an ancestral species of prosimian that had short tarsi? In your answer, be sure to connect what is happening at the molecular (genetic) level to the level of the whole organism." The E-ACORNS rubric is organized around five core concepts, each of which has novice, intermediate, and expert-level statements. Because previous work has shown that student responses to ACORNS prompts vary with context—specifically, whether students are considering the gain or loss of a trait, whether they are analyzing an animal or plant example, and whether they are familiar or unfamiliar with the trait in question (Nehm *et al.*, 2012; Opfer *et al.*, 2012)—we chose two prompts for each survey, one of which referred to a trait gain and one of which referred to a trait loss. In addition, we placed the E-ACORNS prompts in increasing level of difficulty and decreasing level of trait familiarity on the pretest, posttest of Course 1, and posttest of Course 2. This approach controlled for the hypothesis that student performance improved over time because they encountered organisms and traits that are more intuitive when reasoning about

natural selection (Nehm *et al.*, 2012). The scoring rubrics for the E-ACORNS prompts were developed by Sievers *et al.* (2023) and are provided in Supplemental Table S4. It is important to note that both ACORNS and E-ACORNS rubrics record scientific (best evidence) and naïve ideas about how natural selection works. Thus, we analyzed four aspects of this construct: scientific and naïve ideas for both trait-gain and trait-loss scenarios.

For all of these measures of learning, answers to open-response questions were scored by graders who were blind to the source of the data and the goals of the study. In each case, teams of two or three graders were assigned to each construct and trained on the rubric using sample student responses. Each team then followed an iterative process of grading identical questions independently and meeting to reach consensus until interrater reliability scores exceeded 0.80. Once that threshold was exceeded, members of each team scored student responses independently, although we scheduled intermittent group norming sessions to review grading decisions and check for coder drift. Discussions at these norming sessions resulted in near-100% agreement on scoring decisions. Nonsense answers, which occurred in about 2% of total responses, were scored as no response. Additional details on the development and implementation of each rubric are available in Brownell *et al.* (2014), Sievers *et al.* (2023), and Wachtell, L., Gardiner, A., Sievers, M., Dickinson, K., Dy, G.E.C., Glenski, E.H., Mukerji, J., Theobald, E., Tran, E.T., Velasco V., and Freeman, S. (unpublished data). We summed points from each rubric and used these totals in data analysis.

Sample Sizes and Power Analyses

The final data set included only students who had complete demographic information and college entrance examination scores available from the registrar, who had completed the course, and who had submitted both the pre survey in Course 1 and the post survey in Course 2.

To evaluate the concurrent study, we analyzed data from 65 students who completed the CURE and 95 students who did the traditional labs in Course 1 at the same meeting time.

To evaluate student performance in the longitudinal experiment, we hand-scored responses to the open-response prompts from 174 students in the CURE and 226 students in the traditional labs. To analyze potential treatment effects on demographic subgroups that are minoritized in STEM, this sample included all URM and all low-SES students in the CURE data set along with an equal or greater number of non-URM and non–low SES students selected at random. In every case, sample sizes varied among survey questions, as some students left some responses blank. Table 1 presents average sample sizes for the questions disaggregated by demographic groups; exact numbers for each question are given in Supplemental Table S5.

We performed a power analysis to evaluate whether this data set was large enough to detect meaningful differences in how the CURE and traditional labs impacted minoritized students (R Core Team, 2019). To interpret this analysis, we used the guidelines for educational interventions recently proposed by Kraft (2020). Specifically, we regarded an effect size of less than 0.05 as small, 0.05 to less than 0.20 as medium, and 0.20 and above as large (Supplemental Figure S1). For the open-response questions, sample sizes were sufficient to detect large effects on women but not on URM, 1st-gen, or low-SES students. Sample sizes were insufficient to detect small or medium effects in any of the four subgroups analyzed.

TABLE 1. Sample sizes used in analyses of treatment effects on demographic subgroups[a]

	CURE	Traditional
URM	15	66
Non-URM	150	153
Low-SES	29	114
Non–low SES	151	115
Female students	114	152
Male students	66	77
1st-gen	34	75
Continuing generation	141	152

[a]Numbers reported here are averages; precise numbers vary slightly among questions or constructs due to scattered missing responses and are provided in Supplemental Table S5.

Data Analysis

To assess the impact of the CURE on each outcome variable, we designed and evaluated regression models in the R statistical package (R Core Team, 2019). We employed the Akaike information criterion (AIC) during manual backward stepwise model selection on each set of models for each outcome variable until we found the model with the optimal AIC (Theobald *et al.*, 2019). We then used this best model to evaluate the impact of the variable of interest. We visualized results with either box plots of actual values or means and standard errors of fitted values, both from the best models, superimposed on violin plots. Violin plots show the complete data in kernel density plots (smoothed histograms) along the vertical axis, presented symmetrically to support easier interpretation.

For the concurrent experiment, model selection showed that student identity did not need to be included as a random factor in linear regression models that estimated the impact of treatment while controlling for total SAT score as an index of academic preparation and ability. We ran models for each of the four 100-point, 1-hour exams given in the course.

In the longitudinal experiment, we used the total possible scores for each question as the outcome variable except for the "Did you do real research?" data. The scoring rubric for this prompt records whether students answered yes or no and then whether they made positive or negative statements about 15 elements that make research interesting and useful to professional scientists. As a result, we broke the analysis into three: 1) the likelihood of students answering yes, 2) the likelihood that students provided at least one positive warrant, and 3) the likelihood that students gave 0 negative statements.

For each analysis, we began by testing whether student identity should be included as a random factor in the model. Because much of the data represented a bounded set of possible scores, we also used AIC to test whether a censored model was more appropriate. We then estimated the impact of treatment while controlling for pre score and either total SAT score or Course 1 grade as an index of academic preparation and ability. Because we only analyzed data from students who completed both the Course 1 CURE and Course 2 CURE, the outcome variable in

TABLE 2. Effect of treatment on exam scores.

Exam		Estimate	SE	*t* value	*p* value	*n*, *df*
1	Intercept	22.50	1.78	12.6	<<0.001	187, 1181
	Non-lab questions	0.54	0.05	10.0	<<0.001	
	Treatment	−2.56	0.09	−2.7	0.007	
2	Intercept	6.81	1.09	6.3	<<0.001	183, 1181
	Non-lab questions	0.42	0.04	11.2	<<0.001	
3	Intercept	−1.12	1.68	−0.7	0.51	184, 1182
	Non-lab questions	0.38	0.03	11.7	<<0.001	
4	Intercept	8.95	0.54	16.7	<<0.001	173, 1171
	Non-lab questions	0.03	0.01	3.7	<0.001	

the models was the Bio2 post score, with the Bio1 pre score serving as a predictor.

Human Subjects Review

This study was conducted with oversight from the UW Human Subjects Division, application 00003631.

RESULTS

Course Performance

Table 2 summarizes the output of the best regression models on the question of whether treatment predicted differences in exam scores during the concurrent experiment. Students in the traditional labs scored an average of 2.6 points higher on questions related to traditional labs in exam 1(Table 2). This result indicates that, on average, students in the CURE experienced a 2.6% drop in their score on this 100-point test.

There are at least two hypotheses that could explain the difference in exam 1 scores:

1. A traditional lab in which students design their own experiment about trail-following behavior in termites was directly relevant to 29 exam 1 points and could have been especially effective in reinforcing key concepts; and/or
2. A workshop-type lab in which students use pipe cleaners to simulate the stages of meiosis and answer questions about sources of genetic variation was directly relevant to 24 exam 1 points and could have been especially effective.

Unfortunately, we do not have data to address these hypotheses.

Treatment did not, however, appear as a predictor of exam performance in the best models for the following three exams. This means that, all else equal, there were no differences between treatment groups in scores on questions addressed by the traditional labs on exams 2–4.

There are at least two hypotheses that could explain the lack of treatment effect on the final three exams:

1. Relative to exam 1, the smaller number of lab-relevant points available in exams 2–4 made differences between treatment groups undetectable; or
2. The traditional labs did not do an effective job of reinforcing key course concepts, as assessed by the questions posed on exams 2–4.

These results confirm that traditional labs had a measurable effect on student exam performance. When evaluating this benefit of traditional labs, however, it is important to note that the effect size was extremely small: The average difference of 2.6 points represented 0.7% of the total exam points in the course, even though traditional labs were directly relevant to 32% of the total of 400 exam points possible.

Other Measures of Learning

Table 3 summarizes the results of regression models for the aspects of learning that were measured in the longitudinal study. In each case, details on model output are provided as Supplemental Material.

1. **Culture of scientific research.** We analyzed data from three open-response prompts: "What does it mean to think like a scientist?," "What does it mean to do science?," and "Did you do real research in your course lab?" Wachtell, L., Gardiner, A., Sievers, M., Dickinson, K., Dy, G.E.C., Glenski, E.H., Mukerji, J.,Theobald, E., Tran, E.T., Velasco V., and Freeman, S. (unpublished data) have shown that, collectively, the scoring rubrics for these prompts correspond to almost 90% of the culture of scientific research framework developed by Dewey *et al.* (2021). As a result, we interpret student scores on these questions to indicate progress or lack of progress in understanding what it means to be a scientist. The best regression model showed that students in the CURE section did not, on average, provide a larger number of expert-like statements on how scientists think (Supplemental Table S6). In contrast, the models revealed that students in the CURE section had a better understanding of what doing science entails (Figure 1 and Supplemental Table S7) and were more likely to answer "yes" when asked whether they did real research (Figure 2 and Supplemental Table S8a). In response to the "real research" prompt, students were also more likely to give valid warrants to explain why their lab work in the course was real (Supplemental Table S8b) and were less likely to give valid warrants to explain why their lab work in the course was *not* real (Supplemental Table S8c). Examples of positive statements included

   - "It directly contributed to a science project happening at our university."
   - "We tried techniques, some worked and others did not, when they did not, we altered the experiments."
   - "We each performed experiments that have not been done before in these labs and it was real data that no one knew the answers to that we were gathering."

   Examples of negative statements included:

TABLE 3. Summary of results from the longitudinal study

Topic	Outcome variable or construct[a]	Impact of CURE[b]	Model output
Culture of scientific research	What it means to think like a scientist	ns	Supplemental Table S6
	What it means to do science	+	Supplemental Table S7
	Did you do real research in lab?	+	Supplemental Table S8
Experimental design	E-EDAT	ns	Supplemental Table S9
Natural selection	E-ACORNS trait gain	+	Supplemental Table S10a
	E-ACORNS trait-gain misconceptions	ns	Supplemental Table S10b
	E-ACORNS trait loss	+	Supplemental Table S10c
	E-ACORNS trait-loss misconceptions	ns	Supplemental Table S10d

[a]See text for further explanation of each outcome variable.
[b]Plus sign (+) indicates a positive impact on students in the CURE treatment, according to the best model and a p value ≤ 0.05. "ns" indicates no significant impact of the CURE or that treatment was not included in the best model. For details, see the model output tables in the Supplemental Material.

- "Data that was found in these labs were not published or recorded for later deeper analyses."
- "The labs we performed in class answered long known questions and have been done thoroughly before."
- "There was a 'right answer' to many of our lab questions." Taken together, we view these results as strong evidence that compared with students in the traditional labs, CURE students gained a better understanding of what it means to be a scientist.

2. **Experimental design.** The E-EDAT prompts challenge students to design an experiment to test a specific claim; the rubric scores student responses on 17 aspects of experimental design (Sirum *et al.*, 2011; Brownell *et al.*, 2014). Even though our CURE was designed to emphasize the use of controls and other aspects of a rigorous experimental protocol, there was no treatment effect in responses to the E-EDAT prompts. In both treatment groups, scores actually declined over time (Figure 3 and Supplemental Table S9). The lack of treatment effect on the E-EDAT results may result from two conflicting aspects of the CURE's design. The CURE's research question was designed to be experimental in nature—in contrast to the purely descriptive or discovery science emphasized in most CUREs (Hatfull, 2010)—but the CURE tasks themselves were designed to closely mimic a URE, in which participants are primarily responsible for collecting data and rarely, if ever, create the research question and study design. The decline in scores over time that were observed in both treatments may reflect forgetting, as Course 1 class sessions, exams, and labs (both CURE and traditional) had a strong emphasis on experimental design, while Course 2 did not.
3. **Evolution by natural selection.** CURE students scored much better than traditional lab students on the E-ACORNS prompts to explain both the evolution of trait gain and trait loss (Figure 4 and Supplemental Table S10). These positive results could be explained by the work that CURE students did to 1) characterize the molecular basis of antibiotic susceptibility and resistance in bacterial strains and 2) measure changes in the relative fitness of both types of strains. In contrast, there was no treatment effect on the likelihood of stating a misconception about natural selection in response to either the trait-gain prompt or the trait-loss prompt. These negative results should be interpreted cautiously, however, as only about 2% of the responses to E-ACORNS prompts at the end of Course 2 indicated one of the four naïve ideas scored in the rubric. Given this low value, it would be difficult to detect any differences between treatment groups.

FIGURE 1. CURE students may have had a larger increase in understanding what it means to do science. The fitted values shown here control for pre score and SAT total and indicate the average change in mean by treatment ±SE. The p value on the difference in means is 0.055.

DISCUSSION

The results reported here show that, compared with students in traditional labs, students participating in this CURE experienced a small detrimental impact on exam performance on the first exam in Course 1 but gained a more sophisticated understanding of the culture of scientific research and a more expert-like understanding of the connections between genotype, phenotype, and fitness that cause evolution.

Course Performance

The small impact of "losing" seven workshop and guided-inquiry labs—in exchange for CURE labs—on Course 1 exam performance in this study is consistent with recent work showing no gains in conceptual understanding based on lab participation in introductory physics at multiple institutions and an introductory biology course at a single institution (Holmes *et al.*, 2017;

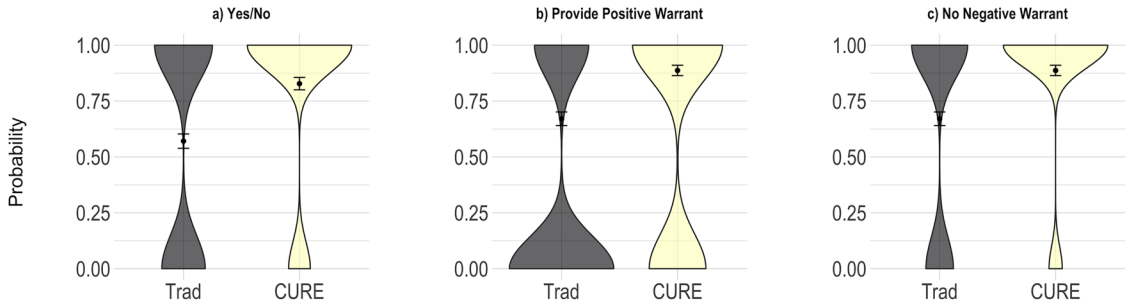

FIGURE 2. Perception of doing real research is higher in CURE labs. Students who did CURE labs were much more likely to respond to the "Did you do real research in lab?" prompt by (a) saying "yes" ($p << 0.0001$), (b) providing at least one positive justification of 15 possible ($p << 0.0001$), and (c) giving no negative reasons ($p << 0.0001$).

Defeo *et al.*, 2020). In addition, a particularly well-designed study at a Hispanic-serving institution randomized students into sections of an introductory biology course with traditional or CURE labs and showed that CURE students did *better* in terms of final course grades (Ing *et al.*, 2021). Because traditional labs are usually designed around the learning goal of reinforcing core course concepts, these results call their value into question. In contrast, we are aware of one study in general chemistry that shows a positive impact of traditional labs in terms of course performance. Matz *et al.* (2012) showed that students who took their general chemistry lab concurrently with lecture did better on course exams compared with students who did not take the lab concurrently. Overall, however, this study adds to a growing literature that challenges the assumption that traditional labs increase conceptual understanding of course content in a meaningful way.

Other Measures of Learning

In terms of measures of learning other than exam performance, our data show strong gains in CURE students' understanding of what Dewey and colleagues (2021) term the culture of scientific research. Those authors developed this construct from an extensive literature review aimed at specifying three defining attributes of scientific culture: 1) practices, or the day-to-day activities of researchers; 2) norms and expectations, meaning the standards that guide behavior in research; and 3) values and beliefs, or the broad ideas that define research as scientific. When Wachtell, L., Gardiner, A., Sievers, M., Dickinson, K., Dy, G.E.C., Glenski, E.H., Mukerji, J., Theobald, E., Tran, E.T., Velasco V., and Freeman, S. (unpublished data; see Supplemental Table S2) developed rubrics to score the open-response prompts used in this study—what it means to think like a scientist, what it means to do science, and whether course labs represented real research—they found that, together, the rubric elements corresponded to 27 of the 31 elements in the culture of scientific research construct. The close correspondence between these rubrics and the Dewey *et al.* (2021) framework allowed us to assess a major goal of CUREs: introducing students to scientific practices and habits of mind (Brownell *et al.*, 2015; Cooper *et al.*, 2019). Our data suggest that participating in this CURE helped students better understand what it means to be a scientist.

Although the lack of treatment effect on scores from the E-EDAT instrument were disappointing, we predict the same

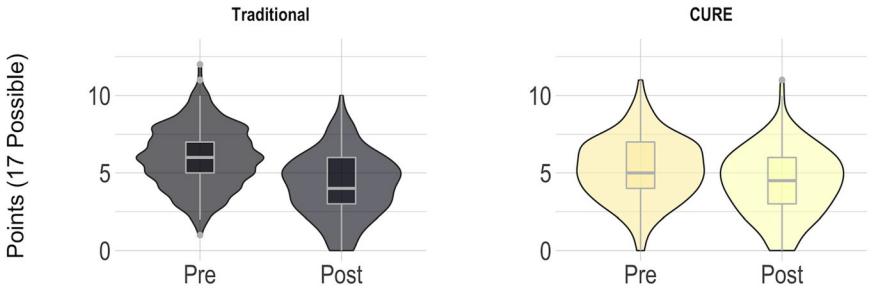

FIGURE 3. Student understanding of experimental design declined in both treatments. The box plots show the medians and interquartile ranges of the actual (not fitted) data. The regression model indicated no treatment effect ($p > 0.05$). The best regression model showed no treatment effect ($p > 0.05$).

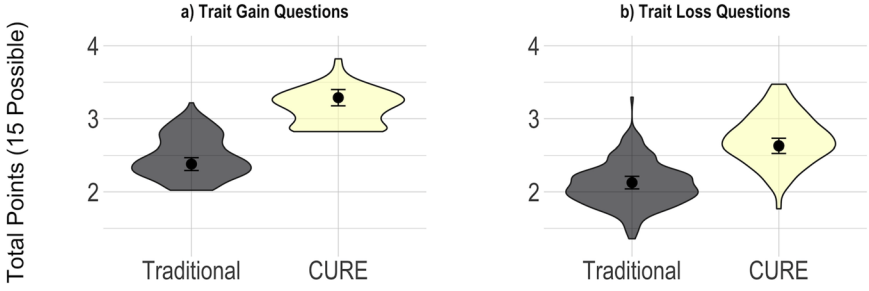

FIGURE 4. Understanding of evolution by natural selection is higher in CURE vs. traditional labs. The fitted values shown in both graphs control for pre score, so they indicate the average change in mean by treatment ±SE. The values in b are also controlled for SAT total score and sex, which were retained in the best model. The best regression models showed treatment effects for (a) trait gain ($p << 0.0001$) and (b) trait loss ($p = 0.0002$).

pattern would be observed in students who are doing a first or possibly even second term of a URE—meaning, when they are not experienced enough to design their own experiments but follow protocols provided by their research mentor. In addition, we predict that inquiry labs that challenge students to design their own experiments, even though the results are not of interest to the broader research community, may be more effective at improving E-EDAT scores than a URE-like CURE such as the one studied here.

Improved scores on the E-ACORNS instrument, which asks students to explain the molecular evolutionary basis of a trait gain or a trait loss, represent an effective answer to calls for introductory biology instruction to focus on the molecular basis of adaptation and the connections between genotypes, phenotypes, and fitness (Smith *et al.*, 2009; Kalinowski *et al.*, 2010; White *et al.*, 2013). These results also make the general point that CUREs can, if designed appropriately, lead to important gains in student understanding of fundamental biological concepts, in addition to promoting research-related ideas.

Impacts on Minoritized Students

Our data do not support the hypothesis that the CURE studied here had a disproportionately large positive impact on underrepresented students, at least for the constructs we measured. This result was disappointing given that 1) promoting retention by underrepresented students is a major goal of CUREs (e.g., Estrada *et al.*, 2016), 2) disproportionate attrition from STEM cannot be reversed unless interventions result in disproportionate benefits for minoritized students, and 3) a power analysis on our sample indicated an ability to detect effective sizes of 0.20—which Lipsey *et al.* (2012) consider the accepted standard for policy change in K–12 education—for women. However, even with our effort to enrich our sample to overrepresent minoritized students, which resulted in 27% of the students in our sample identifying as URM, 37% as 1st-gen, and 54% as low-SES, we would not have been able to detect effect sizes at this level. This observation suggests that, if researchers want to test the hypothesis that CUREs provide disproportionate benefits for minoritized students, effect sizes would need to be extremely high or evidence would need to accumulate from extremely large samples and/or institutions with extraordinarily high percentages of URM, low-SES, or 1st-gen students.

Unfortunately, this study's lack of strong positive signal in terms of disproportionate gains for minoritized students is consistent with the current literature. Although this and other CURE studies are documenting strong gains in learning and attitude for the overall student population, to date, only one study—which may have been impacted by self-selection into the CURE treatment group—has shown a disproportionate benefit for one of the four best-studied minoritized groups: women (Hanauer *et al.*, 2022; for a review of other studies, see Krim *et al.*, 2019). The challenge of designing a CURE that yields disproportionate benefits for other minoritized groups in STEM, and then documenting those benefits rigorously, remains.

Limitations

Because our focus was on student outcomes, we did not measure the impact of the CURE on the research faculty who created the question and may use the data, even though this may be an important benefit of some CUREs (Brownell *et al.*, 2012; Fukami, 2013; Kowalski *et al.*, 2016). For example, CUREs may reduce the tension between investing in teaching or research that plagues tenure-track faculty at research institutions (Fukami, 2013; Shortlidge *et al.*, 2017), and one study suggested that designing and managing a CURE can increase faculty productivity (Kloser *et al.*, 2013). In addition, we have yet to assess the impact on graduate TAs and peer facilitators, although some work has documented benefits to graduate TAs from teaching inquiry versus traditional "cookbook" labs (French and Russell, 2002) or from teaching CUREs (Heim and Holt, 2019). Finally, we did not measure students' persistence in STEM, which is the most important outcome in terms of the original justification for promoting CUREs.

In general, expectations for student outcomes from this and other CUREs need to be tempered by 1) the challenges of having the labs taught by graduate TAs who are not invested in the question or model system (Heim and Holt, 2019) and 2) the limited time on task that occurs. In the CURE studied here, for example, students only spent 23 hours working in the lab. Given that most research groups ask for 5–10 hours of work per week from their undergraduate assistants, time on task in this CURE was equivalent to about 3 weeks in a classical URE. At least some CUREs show increased benefits with increased duration (Shaffer *et al.*, 2014), and the literature on UREs shows that extended duration—often two terms or more—is critical for achieving strong student outcomes in terms of STEM degree attainment, acceptance into graduate programs, and STEM workforce participation (Hernandez *et al.*, 2018). Based on these observations and the results presented here, we hypothesize that efforts to increase the duration and intensity of CUREs may be less important for documenting specific learning gains like understanding the culture of scientific research and how natural selection works and more important for achieving the attitudinal shifts that lead to improved retention in STEM.

Conclusions and Future Work

The data reported here show that a CURE on experimental evolution in *E. coli* produced improvements in understanding of the culture of scientific research and understanding of the molecular basis of adaptation, with an extremely modest trade-off in terms of performance on exam questions related to traditional labs. Introductory course faculty who are considering a change from traditional expository or inquiry labs to CURE labs should be motivated by 1) the small or absent benefits of traditional labs in terms of reinforcing course concepts reported here and elsewhere and 2) the impressive benefits of CUREs in helping students develop interest and expertise in research-related ideas. Motivation for adopting CUREs should be particularly strong in programs that aspire to equip their students for careers in medicine, biotechnology, and basic research.

The student gains documented here are particularly notable for three reasons. First, we collected them during the initial attempts to implement this CURE at scale. After piloting the protocols with groups of 24 students before initiating this study, we increased to four sections with 96 students in the concurrent experiment and then to 25 lab sections and about 600 students in the longitudinal experiment. Second, the longitudinal study compared on-track students in traditional labs versus an off-track cohort in the CURE. Third and most important, the teaching team strictly avoided any messaging to CURE students about

the value of the experience in terms of their intellectual maturation as a scientist, the practical skills they were gaining, their membership in the research community gained through contributing to ongoing work in the Kerr lab, or their ability to use the CURE as a springboard for applying to UREs. Messaging about the importance of the data students were collecting for basic science and clinical medicine was only slightly less limited, focusing only on the concepts of cross-resistance and compensatory mutations and their impacts on therapeutic approaches. One possibility for further work is to add messaging on all of these points reinforced with self-reflection assignments, then document changes in attitudes using published instruments.

Exploring the role of mentoring in student outcomes is another important frontier in CURE research. Mentoring relationships are a primary driver of both positive and negative student responses in UREs (Estrada *et al.*, 2018; Joshi *et al.*, 2019). But if CUREs are offered in large-enrollment introductory courses to fulfill the goal of democratizing access to research experiences, there is an almost inevitable trade-off in terms of how much time faculty and staff can devote to mentoring activities. Designing ways to integrate mentoring from peer facilitators, graduate TAs, and others remains an important challenge for the CURE research community (Grabowski *et al.*, 2008).

Finally, CUREs have attracted intense interest from researchers and policymakers as a tool for solving underrepresentation in STEM, primarily because UREs have played such an important role in retaining URM students (Estrada *et al.*, 2018, Hernandez *et al.*, 2018). As noted, however, few if any CUREs to date have reported the disproportionate benefits that are required to reduce opportunity gaps and mitigate underrepresentation. To document impacts like this and convince students that they belong in STEM—even though they are underrepresented—it may be necessary to design higher-intensity, longer-duration CUREs around scientific questions or societal issues that explicitly impact the communities represented by female, URM, low-SES, and 1st-gen individuals.

ACKNOWLEDGMENTS

We thank members of the Kerr lab for help in designing the experimental question and approach; the UW Biology Education Research Group for helpful discussions on how to study student outcomes; John Parks and the undergraduate peer facilitators and graduate TAs who helped to teach the CURE; and Khoi Ha, Brad Howe, and Melissa Krook for grading student responses. This work was funded by the STEM-Dawgs grant 52008126 from the Howard Hughes Medical Institute, and in part with support by the National Science Foundation under cooperative agreement no. DBI-0939454 (BEACON STC). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

American Association for the Advancement of Science. (2010). *Vision and change in undergraduate biology education: A call to action*. Washington, DC.

Auchincloss, L. C., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., ... & Dolan, E. L. (2014). Assessment of course-based research experiences: A meeting report. *CBE—Life Sciences Education*, *13*, 29–40.

Bangera, G., & Brownell, S. E. (2014). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE—Life Sciences Education*, *13*, 602–606.

Brownell, S. E., Hekmat-Scafe, D. S., Singla, V., Chandler Seawell, P., Conklin Imam, J. F., Eddy, S. L., ... & Cyert, M. S. (2015). A high-enrollment course-based undergraduate research experience improves student conceptions of scientific thinking and ability to interpret data. *CBE—Life Sciences Education*, *14*, 1–14.

Brownell, S. E., & Kloser, M. J. (2015). Toward a conceptual framework for measuring the effectiveness of course-based undergraduate research experiences in undergraduate biology. *Studies in Higher Education*, *40*(3), 525–544. doi: 10.1080/03075079.2015.1004234

Brownell, S. E., Kloser, M. J., Fukami, T., & Shavelson, R. (2012). Comparing the impact of traditionally based "cookbook" and authentic research-based courses on student lab experiences. *Journal of College Science Teaching*, *41*, 36–45.

Brownell, S. E., Wenderoth, M. P., Theobald, R., Okoroafor, N., Koval, M., Freeman, S., ... & Crowe, A. J. (2014). How students think about experimental design: Novel conceptions revealed by in-class activities. *BioScience*, *64*(2), 125–137.

Burnette, J. W. III, & Wessler, S. R. (2013). Transposing from the laboratory to the classroom to generate authentic research experiences for undergraduates. *Genetics*, *193*, 367–375.

Cooper, K. M., Blattman, J. N., Hendrix, T., & Brownell, S. E. (2019). The impact of broadly relevant novel discoveries on student project ownership in a traditional lab course turned CURE. *CBE—Life Sciences Education*, *18*, ar57.

Cooper, K. M., Soneral, P. A., & Brownell, S. E. (2017). Define your goals before you design a CURE: A call to use backward design in planning course-based undergraduate research experiences. *Journal of Microbiology & Biology Education*, *18*(2), 28656069, doi: 10.1128/jmbe/v18i2/1287

Corwin, L. A., Runyon, C. R., Ghanem, E., Sandy, M., Clark, G., Palmer, G. C., ... & Dolan, E. L. (2018). Effects of discovery, iteration, and collaboration in laboratory courses on undergraduates' research career intentions fully mediated by student ownership. *CBE—Life Sciences Education*, *17*(2), ar20.

Corwin, L. A., Runyon, C., Robinson, A., & Dolan, E. L. (2015). The Laboratory Course Assessment Survey: A tool to measure three dimensions of research-course design. *CBE—Life Sciences Education*, *14*(4), ar37.

DeFoe, D. J., Bibler, A., & Gerken, S. (2020). The effect of a paired lab on course completion and grades in nonmajors introductory biology. *CBE—Life Sciences Education*, *19*, ar36, 1–12.

Dewey, J., Roehrig, G., & Schuchardt, A. (2021). Development of a framework for the culture of scientific research. *CBE—Life Sciences Education*, *20*, ar65, 1–17.

Dickinson, K., Mukerji, J., Graham, S., Warfield, L., & Kerr, B. (2021). A course-based undergraduate research experience on the evolution of antibiotic resistance and its molecular basis. *Preprints*, doi: 10.20944/preprints202107.0420.v1

Elgin, S. C. R., Bangera, G., Decatur, S. M., Dolan, E. L., Guertin, L., Newstetter, W. C., ... & Labov, J. B. (2016). Insights from a convocation: Integrating discovery-based research into the undergraduate curriculum. *CBE—Life Sciences Education*, *15*, 1–7.

Estrada, M., Burnett, M., Campbell, A. G., Campbell, P. B., Denetclaw, W. F., Gutiérrez, ... & Zavala, M. (2016). Improving underrepresented minority persistence in STEM. *CBE—Life Sciences Education*, *15*, es5, 1–10.

Estrada, M., Hernandez, P. R., & Schultz, P. W. (2018). A longitudinal study of how quality mentorship and research opportunities integrate underrepresented minorities into STEM careers. *CBE—Life Sciences Education*, *17*, ar9, 1–13.

Freeman, S., Okoroafor, N. O., Gast, C. M., Koval, M., Nowowiejski, D., O'Connor, E., ... & Fang, F. C. (2016). Crowdsourced data indicate widespread multidrug resistance in skin flora of healthy young adults. *Journal of Microbiology and Biology Education*, *17*(1), 172–182.

French, D., & Russell, C. (2002). Do graduate teaching assistants benefit from teaching inquiry-based laboratories? *BioScience*, *52*, 1036–1041.

Fukami, T. (2013). Integrating inquiry-based teaching with faculty research *Science*, *339*, 1536–1537.

Gasper, B. J., & Gardner, S. M. (2013). Engaging students in authentic microbiology research in an introductory biology course is correlated with

gains in student understanding of the nature of authentic research and critical thinking. *Journal of Microbiology and Biology Education*, *14*, 25–34.

Gentile, J., Brenner, K., & Stephens, A. (2017). *Undergraduate research experiences for undergraduates: Successes, challenges, and opportunities*. Washington, DC: National Academies Press.

Grabowski, J. J., Heely, M. E., & Brindley, J. A. (2008). Scaffolding faculty-mentored research experiences for first-year students. *Council on Undergraduate Research Quarterly*, *29*, 41–47.

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, *332*, 1213–1216.

Hanauer, D. I., & Dolan, E. L. (2014). The Project Ownership Survey: Measuring differences in scientific inquiry experiences. *CBE—Life Sciences Education*, *13*, 149–158. doi 10.1187/cbe.13-06-0123

Hanauer, D. I., Frederick, J., Fotinakes, B., & Strobel, S. A. (2012). Linguistic analysis of project ownership for undergraduate research experiences. *CBE—Life Sciences Education*, *11*(4), 378–385.

Hanauer, D. I., Graham, M. J., Jacobs-Sera, D., Garlena, R. A., Russell, D. A., Sivanathan, V., … & Hatfull, G. F. (2022). Broadening access to STEM through the community college: Investigating the role of course-based research experiences (CREs). *CBE—Life Sciences Education*, *21*, ar38, 1–16.

Hanauer, D. I., Graham, M. J., SEA-PHAGES Betancur, L., Bobrownicki, A., Cresawn, S. G., … & Hatfull, G. F. (2017). An inclusive research education community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning. *Proceedings of the National Academy of Sciences USA*, *114*(51), 13531–13536.

Harris, R. B., Mack, M. R., Bryant, J., Theobald, E. J., & Freeman, S. (2020). Reducing achievement gaps in undergraduate general chemistry could lift underrepresented students into a "hyperpersistent zone." *Science Advances*, *6*, eaaz5687.

Harris, S. E., & Bellino, M. (2013). DNA barcoding from NYC to Belize. *Science*, *342*, 1462–1463.

Harrison, M., Dunbar, D., Ratmansky, L., Boyd, K., & Lopatto, D. (2011). Classroom-based science research at the introductory level: Changes in career choices and attitude. *CBE—Life Sciences Education*, *10*, 279–286.

Hatfull, G. E. (2010). Bacteriophage research: Gateway to learning sciences. *Microbe*, *5*, 243–250.

Heim, A. B., & Holt, E. A. (2019). Benefits and challenges of instructing introductory biology course-based undergraduate research experiences (CUREs) as perceived by graduate teaching assistants. *CBE—Life Sciences Education*, *18*, ar43, 1–12.

Hernandez, P. R., Woodcock, A., Estrada, M., & Schultz, P. W. (2018). Undergraduate research experiences broaden diversity in the scientific workforce. *BioScience*, *68*(3), 204–211.

Herrera, F. A., & Hurtado, S. (2011). *Maintaining initial interests: Developing science, technology, engineering, and mathematics (STEM) career aspirations among underrepresented racial minority students*. Los Angeles: UCLA School of Education and Information Studies.

Holmes, N. G., Olsen, J., Thomas, J. L., & Wieman, C. E. (2017). Value added or misattributed? A multi-institution study on the educational benefit of labs for reinforcing physics content. *Physical Review Physics Education Research*, *13*, 010129.

Indorf, J. L., Weremijewicz, J., Janos, D. P., & Gaines, M. S. (2019). Adding authenticity to inquiry in a first-year, research-based, biology laboratory course. *CBE—Life Sciences Education*, *18*, ar38, 1–15.

Ing, M., Burnette, J. M. III, Azzam, T., & Wessler, S. R. (2021). Participation in a course-based undergraduate research experience results in higher grades in the companion lecture course. *Educational Researcher*, *50*(4), 205–214.

Irby, S. M., Pelaez, N. J., & Anderson, T. R. (2018). How to identify the research abilities that instructors anticipate students will develop in a biochemistry course-based undergraduate research experience (CURE). *CBE—Life Sciences Education*, *17*, es4, 1–14.

Jacob, N. P. (2012). Investigating Arabia Mountain: A molecular approach. *Science*, *335*, 1588–1589.

Jordan, T. C., Burnett, S. H., Carson, S., Caruso, S. M., Clase, K., DeJong, R. J., … & Hatfull, G. F. (2014). A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. *mBio*, *5*(1), e01051–13.

Joshi, M., Aikens, M. L., & Dolan, E. L. (2019). Direct ties to a faculty mentor related to positive outcomes for undergraduate researchers. *BioScience*, *69*(5), 389–397.

Kalinowski, S. T., Leonard, M. J., & Andrews, T. M. (2010). Nothing in evolution makes sense except in the light of DNA. *CBE—Life Sciences Education*, *9*, 87–97.

Kloser, M. J., Brownell, S. E., Shavelson, R. J., & Fukami, T. (2013). Effects of a research-based ecology lab course: A study of nonvolunteer achievement, self-confidence, and perception of lab course purpose. *Journal of College Science Teaching*, *42*, 72–81.

Kowalski, J. R., Hoops, G. C., & Johnson, R. J. (2016). Implementation of a collaborative series of classroom-based undergraduate research experiences spanning chemical biology, biochemistry, and neurobiology. *CBE—Life Sciences Education*, *15*, ar55, 1–17.

Kraft, M. A. (2020). Interpreting effect sizes of educational interventions. *Educational Researcher*, *49*(4), 241–253.

Krim, J. S., Coté, L. E., Schwartz, R. S., Stone, E. M., Cleeves, J. J., Barry, K. J., … & Rebar, B. M. (2019). Models and impacts of science research experiences: A review of the literature of CUREs, UREs, and TREs. *CBE—Life Sciences Education*, *18*, ar65, 1–14.

Lau, J. M., & Robinson, D. L. (2009). Effectiveness of a cloning and sequencing exercise on student learning with subsequent publication in the National Center for Biotechnology Information GenBank. *CBE—Life Sciences Education*, *8*, 326–337.

Lipsey, M. W., Puzio, K., Yun, C., Hebert., M. A., Steinka-Fry, K., Cole, M. W., … & Busick, M. D. (2012). *Translat*ing *the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Services, U.S. Department of Education.

Luckie, D. B., Maleszewski, J. J., Loznak, S. D., & Krha, M. (2004). Infusion of collaborative inquiry throughout a biology curriculum increases student learning: A four-year study of "Teams and Streams." *Advances in Physiology Education*, *287*, 199–209.

Matz, R. L., Rothman, E. D., Krajcik, J. S., & Holl, M. M. B. (2012). Concurrent enrollment in lecture and laboratory enhances student performance and retention. *Journal of Research in Science Teaching*, *49*, 659–682.

McHugh, P. P. (2016). The impact of compensation, supervision and work design on internship efficacy: Implications for educators, employers and prospective interns. *Journal of Education and Work*, *30*(4), 367–382. doi: 10.1080/13639080.2016.1181729

National Academies of Sciences, Engineering, and Medicine. (2016). *Barriers and opportunities for 2-year and 4-year STEM degrees*. Washington, DC: National Academies Press.

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *American Biology Teacher*, *74*, 92–96.

Olimpo, J. T., Fisher, G. R., & DeChenne-Peters, S. E. (2016). Development and evaluation of the *Tigriopus* course-based undergraduate research experience: Impacts on students' content knowledge, attitudes, and motivation in a majors introductory biology course. *CBE—Life Sciences Education*, *15*, ar71, 1–15.

Opfer, J. E., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: Knowing what students know about evolution. *Journal of Research in Science Education*, *49*(6), 744–777.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rodenbusch, S. E., Hernandez, P. R., Simmons, S. L., & Dolan, E. L. (2016). Early engagement in course-based research increases graduation rates and completion of science, engineering, and mathematics degrees. *CBE—Life Sciences Education*, *15*, ar20, 1–10.

Rodrigo-Pieris, T., Xiang, L., & Cassone, V. M. (2018). A low-intensity, hybrid design between a "traditional" and a "course-based" research experience yields positive outcomes for science undergraduate freshmen and shows potential for large-scale application. *CBE—Life Sciences Education*, *17*, ar53, 1–18.

Shaffer, C. D., Alvarez, C., Bailey, C., Barnard, D., Bhalla, S., Chandrasekaran, C., … & Elgin, S. C. R. (2010). The Genomics Education Partnership: Successful integration of research into laboratory classes at a diverse group of undergraduate institutions. *CBE—Life Sciences Education*, *9*, 55–69.

Shaffer, C. D., Alvarez, C. J., Bednarski, A. E., Dunbar, D., Goodman, A. L., Reinke, C., ... & Elgin, S. C. (2014). A course-based research experience: How benefits change with increased investment in instructional time. *CBE—Life Sciences Education*, *13*(1), 111–130.

Shortlidge, E. E., Bangera, G., & Brownell, S. E. (2017). Each to their own CURE: Faculty who teach course-based undergraduate research experiences report why you too should teach a CURE. *Journal of Microbiology & Biology Education*, *18*, 1260.

Sievers, M., Reemts, C., Dickinson, K., Barreras, I. B., Mukerji, J., Theobald, E. J., ... & Freeman, S. (2023). Assessing how well students understand the molecular basis of evolution by natural selection. *Biochemistry and Molecular Biology Education*. https//doi.org.10.1002/bmb.21697

Sirum, K., & Humburg, J. (2011). The Experimental Design Ability Test (EDAT). *Bioscene: Journal of College Biology Teaching*, *37*, 8–16.

Smith, J. J., Baum, D. A., & Moore, A. (2009). The need for molecular genetic perspectives in evolutionary education (and vice versa). *Trends in Genetics*, *25*, 427–429.

Stanich, C. A., Pelch, M. A., Theobald, E. J., & Freeman, S. (2018). A new approach to supplementary instruction narrows achievement and affect gaps for underrepresented minorities, first-generation students, and women. *Chemistry Education Research and Practice*, *19*, 846–866.

Tekin, E., White, C., Kang, T. M., Singh, N., Cruz-Loya, M., Damoiseaux, R., ... & Yeh, P. J. (2018). Prevalence and patterns of higher-order drug interactions in *Escherichia coli*. *npj Systems Biology and Applications*, *4*, 31; doi: 10.1038/s41540-018-0069-9

Theobald, E. J., Aikens, M., Eddy, S., & Jordt, H. (2019). Beyond linear regression: A reference for analyzing common data types in discipline based education research. *Physical Review Physics Education Research*, *15*, 020110.

Wang, J. T. H., Daly, J. N., Willner, D. L., Patil, J., Hall, R. A., Schembri, M. A., ... & Hugenholtz, P. (2015). Do you kiss your mother with that mouth? An authentic large-scale undergraduate research experience in mapping the human oral microbiome. *Journal of Microbiology and Biology Education*, *16*, 50–60.

Wei, C. A., & Woodin, T. (2011). Undergraduate research experiences in biology: Alternatives to the apprenticeship model. *CBE—Life Sciences Education*, *10*, 123–131.

White, P. J. T., Heidemann, M., Loh, M., & Smith, J. J. (2013). Integrative cases for teaching evolution. *Evolution: Education and Outreach*, *6*, 17–23.

Wiley, E. A., & Stover, N. A. (2014). Immediate dissemination of student discoveries to a model organism database enhances classroom-based research experiences. *CBE—Life Sciences Education*, *13*(1), 131–138.

Wilson, A. E., Pollock, J. L., Billick, I., Domingo, C., Fernandez-Figueroa, E. G., Nagy, E.S., ... & Summers, A. (2018). Assessing science training programs: structured undergraduate research programs make a difference. Bioscience, *68*, 529–534.

Wischusen, S. M., & Wischusen, E. W. (2007). Biology Intensive Orientation for Students (BIOS): A biology "boot camp." *CBE—Life Sciences Education*, *6*, 172–178.