Randomized Controlled Trial of a Cohesive Eight-Week Evolution Unit That Incorporates Molecular Genetics and Principles of the *Next Generation Science Standards*

D. Drits-Esser,[†]* J. Hardcastle,[‡] K. M. Bass,[§] S. Homburger,[†] M. Malone,[†] K. Pompei,[†] G. E. DeBoer,[‡] and L. A. Stark[†]

[†]Genetic Science Learning Center, University of Utah, Salt Lake City, Utah, 84102; [‡]AAAS Project 2061, Washington, DC 20005; [§]Rockman et al Cooperative, Inc., Berkeley, CA 94703

ABSTRACT

In response to calls for curricular materials that integrate molecular genetics and evolution and adhere to the K–12 *Next Generation Science Standards* (NGSS), the Genetic Science Learning Center (GSLC) at the University of Utah has developed and tested the "Evolution: DNA and the Unity of Life" curricular unit for high school biology. The free, 8-week unit illuminates the underlying role of molecular genetics in evolution while providing scaffolded opportunities to engage in making arguments from evidence and analyzing and interpreting data. We used a randomized controlled trial design to compare student learning when using the new unit with a condition in which teachers used their typical (NGSS-friendly) units with no molecular genetics. Results from nationwide testing with 38 teachers (19 per condition) and their 2269 students revealed that students who used the GSLC curriculum had significantly greater pre/post gain scores in their understanding of evolution than students in the comparison condition; the effect size was moderate. Further, teacher implementation data suggest that students in the treatment condition had more opportunities to engage in argumentation from evidence and have in-class discussions than students in the comparison classes. We consider study implications for the secondary and postsecondary science education community.

INTRODUCTION

Evolution is a unifying, fundamental topic for secondary and postsecondary science students. "The core ideas in the life sciences culminate with the principle that evolution can explain how the diversity that is observed within species has led to the diversity of life across species" (National Research Council [NRC], 2012, p. 140). However, K–12 and postsecondary students' difficulty understanding evolution is well documented (e.g., Gregory, 2009; Flanagan and Roseman, 2011; Borgerding *et al*., 2015).

One promising area of research in evolution education involves teaching and learning strategies that provide opportunities for students to think "across organizational levels." This includes distinguishing between different levels of organization, interrelating concepts at different levels of organization, and being able to think back and forth between these levels (reviewed in Jördens *et al*., 2016). While thinking across levels has been identified as necessary for understanding evolutionary principles, research has shown it to be challenging for students (e.g., Ferrari and Chi, 1998). Engaging students in thinking across levels, particularly at the level of genetic variation by grounding evolutionary phenomena in genetics concepts, has been shown to help in evolution understanding (reviewed in Jördens *et al*., 2016). Using ideas from *molecular* genetics, in which the focus is on variation at the level of DNA and protein

Tessa C. Andrews, *Monitoring Editor*
Submitted Jan 17, 2020; Revised Jun 28, 2021; Accepted Jul 8, 2021
CBE Life Sci Educ September 1, 2021 20:ar50
DOI:10.1187/cbe.20-01-0008

*Address correspondence to: D. Drits-Esser (dina.drits@utah.edu).

sequences rather than at the trait level, has been shown to foster deeper learning of evolution topics (e.g., Kalinowski *et al*., 2010; White *et al*., 2013; Jördens *et al*., 2016). Student learning also showed significant improvement when genetics was taught before—rather than after—evolution, particularly for low-performing students (Mead *et al.*, 2017).

Further, K–12 researchers have called for evolution instruction that uses the principles of the *Next Generation Science Standards* (NGSS; NGSS Lead States, 2013), based on the seminal document *A Framework for K–12 Science Education* (NRC, 2012). Enacting the vision in these documents promotes science teaching and learning that foster deep, integrated student understanding of science and scientific processes. In addition, it promotes an understanding of connections across the disciplines of science and the integration of science practices into science instruction. Both documents advocate for instruction that includes the practices of analyzing and interpreting data and arguing from evidence. Studies have reported student content knowledge gains when argumentation is taught and practiced explicitly in the classroom (Bell and Linn, 2000; Zohar and Nemet, 2002; Asterhan and Schwarz, 2007) and when students are provided opportunities to analyze and interpret data (e.g., Bray Speth *et al*., 2009; Beardsley *et al*., 2011).

Instructors require high-quality curricular materials aligned with NGSS to achieve student learning in evolution (NRC, 2012). Curricula that provide students with opportunities to make relevant scientific conceptual connections and to use science practices lead to positive student learning outcomes (Glaze and Goldston, 2015). However, few educational resources, textbooks, or other curricular materials at the high school and undergraduate levels integrate NGSS science practices and the content areas of both evolution and genetics. With regard to integrating evolution and genetics, existing materials often present these topics as disparate (Kalinowski *et al.*, 2010; White *et al.*, 2013). This creates some uncertainty for educators in how to structure their courses for optimal student learning and leaves it to educators to make the connections in their classes. White *et al*. (2013) argued that, without integrated frameworks or perspectives, student learning is inherently compartmentalized and disconnected, preventing the development of a deep understanding of evolution. To meet this need, a call has been issued for development of additional, integrated curricular resources and for empirical data to explore this relationship and its effectiveness for learning (e.g., Jördens *et al*., 2016). In response, we are investigating ways to help educators leverage vertical thinking while adhering to the principles of NGSS through the development of new curricular materials.

The study reported here addresses these gaps in evolution curricula. The Genetic Science Learning Center (GSLC) at the University of Utah has developed "Evolution: DNA and the Unity of Life," a freely available 8-week replacement unit for high school biology that uses the principles of NGSS by incorporating the science practices of arguing from evidence and analyzing and interpreting data while integrating molecular genetics and evolution throughout the unit. The unit maps each evolution learning goal directly back to DNA (e.g., natural selection acts on heritable variation in DNA that confers a reproductive advantage, causing a change in the characteristics of populations over time). Because thinking across levels is challenging for students (Ferrari and Chi, 1998), the GSLC used research-based learning progressions and relevant learning theories to inform this conceptual integration (for a full description of the theoretical and conceptual framing, see Homburger *et al.*, 2019). This study was conducted as the final efficacy phase of a multiyear curricular evaluation process.

In this research, we examine differences in student learning outcomes and teacher-reported implementation of evolution curricular units at the high school level. This study was conducted using a randomized controlled (RCT) study design involving 38 teachers from across the United States and their 2269 students. To our knowledge, the efficacy of high school evolution units has not been studied using this type of design with large participant numbers. We compared the new unit with a comparison condition composed of teachers' regular evolution units that used the same NGSS science content standards as the treatment condition but did *not* include molecular genetics topics; both conditions used NGSS science practices. We use these data to offer insights into the differences between the conditions and the factors that may have contributed to the assessment and implementation results.

Research Questions

The research studies summarized earlier show the importance of teaching across organizational levels to provide a context for student learning of evolution concepts. In this study, we sought to add to this literature by conducting a national RCT that compares a cohesive, integrated evolution unit with teachers' typical evolution units. We hypothesized that a cohesive evolution unit that uses the principles of NGSS and integrates across biological organizational levels will foster better learning outcomes and teacher experiences than teachers' typical curricula (business as usual).

We based our investigation around two research questions: *First*, what is the difference in student learning outcomes between two curricular conditions: 1) the treatment condition, which is a cohesive evolution unit that integrates NGSS evolution and molecular genetics science content and relevant science practices, particularly argumentation; and 2) the comparison condition, which is an aggregate of teachers' typical evolution units that incorporate the same NGSS evolution science content, include science practices, but do not use molecular genetics concepts to teach evolution? Because one area of interest was the impact on student learning of incorporating molecular genetics in the study of evolution, it would have been preferable to compare the new unit with and without the molecular genetics content (or to have a three-armed study comparing these conditions with a unit without molecular genetics). However, removing molecular genetics from the unit was not possible, as these topics are so closely integrated throughout the entire unit.

Second, what did implementation of curricula look like across the two conditions, and what were the key similarities and differences? We unpack and compare curricular features from both conditions to speculate on what may have contributed to student learning and to teacher implementation and satisfaction.

This work provides high school biology educators with efficacy data for a cohesive replacement evolution and molecular genetics curricular unit that incorporates the principles of NGSS. It also offers biology education writ large empirical data around the question of which curricular elements may contribute to student learning of evolution. We discuss study implications for secondary and postsecondary curriculum developers, researchers, and educators.

METHODS—UNIT EFFICACY TESTING

Research Design

This classroom field test used an RCT design to compare pre/post evolution learning gains between students whose teachers were randomly assigned, by an external evaluator, to one of two conditions. The treatment condition (the new unit) integrated evolution and molecular genetics concepts, while the comparison condition used evolution-only concepts (teachers' regular evolution units). Comparison teachers were asked to refrain from teaching molecular genetics that year until after the field test. Teachers in both conditions could use classical genetics concepts (trait-level ideas) before the field test; it was the integration of the molecular genetics concepts that we were examining in this research.

Similarities between the Treatment and Comparison Units. The units in both conditions were required to adhere to the principles of the national K–12 science standards, the *Next Generation Science Standards* (NGSS Lead States, 2013). The NGSS promote science teaching and learning that foster deep, integrated student understanding of science and scientific processes. NGSS-aligned classrooms include opportunities for students to learn and practice each of three dimensions of learning: *disciplinary core ideas* (key ideas in science), *science practices* (what scientists do to investigate the natural world), and *crosscutting concepts* (making connections across the disciplines of science). Additionally, NGSS curricular materials must articulate investigable questions that address core ideas and must build instruction and student tasks around real-world phenomena (NRC, 2015).

Teachers in both conditions used the same NGSS disciplinary core ideas for biological evolution, which included:

Evidence of Common Ancestry and Diversity (HS-LS4.A)

- Genetic information provides evidence of evolution. DNA sequences vary among species, but there are many overlaps; in fact, the ongoing branching that produces multiple lines of descent can be inferred by comparing the DNA sequences of different organisms. Such information is also derivable from the similarities and differences in amino acid sequences and from anatomical and embryological evidence.

Natural Selection (HS-LS4.B)

- Natural selection occurs only if there are both 1) variation in the genetic information between organisms in a population and 2) variation in the expression of that genetic information—that is, trait variation—that leads to differences in performance among individuals.
- The traits that positively affect survival are more likely to be reproduced, and thus are more common in the population.

Adaptation (HS-LS4.C)

- Evolution is a consequence of the interaction of four factors: 1) the potential for a species to increase in number, 2) the genetic variation of individuals in a species due to mutation and sexual reproduction, 3) competition for an environment's limited supply of the resources that individuals need to survive and reproduce, and 4) the ensuing proliferation of those organisms that are better able to survive and reproduce in that environment.
- Natural selection leads to adaptation, that is, to a population dominated by organisms that are anatomically, behaviorally, and physiologically well suited to survive and reproduce in a specific environment. That is, the differential survival and reproduction of organisms in a population that have an advantageous heritable trait leads to an increase in the proportion of individuals in future generations that have the trait and to a decrease in the proportion of individuals that do not.
- Adaptation also means that the distribution of traits in a population can change when conditions change.
- Changes in the physical environment, whether naturally occurring or human induced, have thus contributed to the expansion of some species, the emergence of new distinct species as populations diverge under different conditions, and the decline—and sometimes the extinction—of some species.

The science practices and crosscutting concepts that are used in the treatment condition are discussed in the next section. These curricular components varied across the units in the comparison conditions and will be discussed in the *Results* section.

The Treatment Unit. The new unit, "Evolution: DNA and the Unity of Life" (GSLC, 2018a,b), is a free, 8-week replacement curricular unit for introductory high school biology students. It includes both technology-based and paper-based materials for teachers and students. Technology is used to introduce phenomena and engage students' interest in them, to depict dynamic or molecular-level processes, to address misconceptions, and to provide summaries of key concepts. In some cases, it is used to enable students to explore topics in depth in a less time-consuming way. Paper-based materials support students in data analysis and sense-making and in organizing and connecting concepts and provide manipulatable models.

The GSLC's evolution unit integrates molecular genetics throughout, building from simpler to more complex levels of understanding. Students analyze data that connect molecular and evolution ideas. Scaffolded practice builds students' ability to develop arguments based on molecular evidence. The GSLC's goal was to provide students with a rich understanding of evolution through a unit that is coherent (Fortus and Krajcik, 2012; NGSS Lead States, 2013) in its content and science practices, learning goals and objectives, assessments, language, visuals, and teacher supports. While molecular genetics and evolution are integrated throughout the 8-week unit, the evolution content comprises approximately 5 weeks.

In developing the unit, the GSLC unpacked the relevant evolution and heredity/genetics *disciplinary core ideas* and grouped the resulting components into topics that support the integration of these content areas; the evolution ideas are listed in the previous section. The GSLC incorporated only the components of the NGSS genetics ideas necessary for explaining the molecular concepts that underlie evolution and that were useful in creating a common "molecular genetics language" through which to convey evolution concepts. These NGSS ideas include:

Structure and Function (HS-LS1.A)

- All cells contain genetic information in the form of DNA molecules. Genes are regions in the DNA that contain the instructions that code for the formation of proteins, which carry out most of the work of cells.

Inheritance of Traits (HS-LS3.A)

- Each chromosome consists of a single very long DNA molecule, and each gene on the chromosome is a particular segment of that DNA. The instructions for forming species' characteristics are carried in DNA.

Variation of Traits (HS-LS3.B)

- In sexual reproduction, chromosomes can sometimes swap sections during the process of meiosis (cell division), thereby creating new genetic combinations and thus more genetic variation. Although DNA replication is tightly regulated and remarkably accurate, errors do occur and result in mutations, which are also a source of genetic variation.

Five modules emerged from the GSLC's grouping of evolution and genetics ideas, with the enduring understanding for each mapping directly back to DNA:

- *Shared Biochemistry of Life: DNA* and the proteins it encodes shape the characteristics of all living things.
- *Common Ancestry: DNA* underlies all evidence for common ancestry.
- *Heredity: DNA* mutations and allele shuffling during sexual reproduction give rise to variation in populations.
- *Natural Selection:* Natural selection acts on heritable variation in *DNA* that confers a reproductive advantage, causing a change in the characteristics of populations over time.
- *Speciation:* Mutation (*DNA*), allele shuffling (*DNA*), and natural selection acting on multiple traits over many generations in reproductively isolated populations cause the divergence in characteristics of living things.

The GSLC also incorporated into the unit what it determined to be the most relevant *science practices* and *crosscutting concepts*. As described earlier, the science practices of making arguments from evidence (Bell and Linn, 2000; Zohar and Nemet, 2002; Asterhan and Schwarz, 2007) and analyzing and interpreting data (e.g., Bray Speth *et al.*, 2009; Beardsley *et al.*, 2011) have been shown to increase students' evolution content understanding. The GSLC therefore selected these as the unit's target practices. Specifically, the unit provides students with opportunities to analyze and interpret skill level–appropriate data from published scientific research about phenomena, and it engages students in the construction of evidence-based arguments based on these phenomena. The crosscutting concepts of patterns and of cause and effect fit into discussions across both evolution and genetics.

See Homburger *et al.* (2019) for a comprehensive description of the unit, its theoretical underpinnings and design principles, and its development process. That work also reports on the unit's feasibility and usability for teachers and on results from a national pilot test with students in grades 9 and 10 introductory biology that found significant gains in students' learning and skill in identifying claims, evidence, and reasoning in scientific arguments. The unit can be accessed at:

- Teacher site: http://teach.genetics.utah.edu/content/evolution
- Student site: http://learn.genetics.utah.edu/content/evolution

The Comparison Units. Comparison teachers used their regular NGSS-type evolution units/curricula, but they did *not* provide students with instruction on the NGSS disciplinary core ideas of genetic inheritance and genetic variation until after the field test was complete. Teachers' applications to participate in the study showed that their evolution units were between 3 and 6 weeks (one of the selection criteria, discussed in more detail later, was trying to match the approximately 5 weeks of evolution in the treatment unit). Eighty-four percent of the teachers used textbooks for teaching evolution topics. Eighty-four percent typically taught genetics (molecular or Mendelian) before evolution, while the rest (16%) typically taught evolution before genetics. The science practices and crosscutting concepts used in teachers' regular units varied among the different units. We describe the comparison units, including their use of science practices and crosscutting concepts, in greater detail in the *Results* section.

RCT Procedures

Participation Inclusion Criteria. We recruited 46 high school teachers (22 treatment, 24 comparison) via the GSLC's email list of more than 20,000 educators nationwide. Participants represented 23 states and diverse teaching contexts, diverse student demographics (socioeconomic, linguistic, ethnic, and racial diversity), and teaching schedules (alternating vs. daily). Teachers were included in the study if they met the following criteria:

- Teaching in an NGSS-oriented classroom, school, and/or district
- Teaching at least two introductory or honors biology (grades 9 or 10) sections
- Willingness to participate regardless of assignment to treatment or comparison condition
- Ability to access Internet-enabled computers or tablets, one per student, for pre/posttesting and portions of the unit
- Willingness to teach their heredity/genetics unit after evolution, if assigned to the comparison condition

Permissions and Randomization. The project's external evaluator assigned teachers randomly to a condition, and we notified teachers about their assignments by email. We secured written permission from principals and, when applicable, support from the district's external research committee. We also received university Institutional Review Board and school/district approval before conducting the research. Communications between the teachers and project staff occurred regularly throughout the year.

Field Test Procedures. During the summer before field-testing, teachers in each condition took part in a condition-specific webinar that included training on the RCT procedures. During their webinar, teachers assigned to the treatment condition also received brief training on how to use the unit materials in their classrooms. Teachers then proceeded with the field test during the academic year. Before beginning their units, all teachers administered an online student content knowledge pretest. Next, teachers in the treatment condition taught the entire new unit in each of their biology sections with no external materials. Teachers in the comparison condition taught their regular evolution units in each of their biology sections. While they were teaching their units, all teachers were asked to complete online logs detailing their daily teaching activities (see *Teacher Implementation Data* section for more details). At the conclusion of their respective units, all teachers administered an online student content knowledge posttest and completed an end-of-unit survey in which they were asked to provide information about their experiences with the unit and their perceived impact of the unit on themselves and their students (see *Teacher Implementation Data* section for more details). When all steps were complete, teachers were compensated for their participation.

Data Inclusion Criteria. The results represent data from 38 teachers (n = 19 treatment, n = 19 comparison) who completed the research requirements and remained in the analytic sample and their 2269 students (n = 1165 treatment, n = 1104 comparison) who completed both the pre- and posttests. Student demographics were 50% female and 50% male or not indicated, 9% English not primary language, 36% free or reduced-price lunch, and 36% underrepresented ethnic or racial groups. There were no significant differences in demographic categories between conditions. See Appendix A in the Supplemental Material for a summary of student demographics for each teacher and Appendix B in the Supplemental Material for a summary of demographic statistics for treatment and comparison groups. The overall study attrition was 17%, with 14% attrition from the treatment condition and 21% attrition from the comparison condition. The differential attrition rate of 7% compared with the overall attrition of 17% meets the optimistic boundary for attrition by the What Works Clearinghouse (WWC, 2017). See Appendix C in the Supplemental Material for numbers of teachers and students in each condition and the differential attrition rates.

Student Assessment

Assessment Development. Student pre/postassessment items were developed by AAAS Project 2061, which nationally pilot-tested and revised the items according to established procedures (NRC, 2014). The assessment team began item construction by reviewing NGSS standards and identifying the evolution constructs to be measured. Next, they identified relevant student misconceptions by reviewing published literature and interviewing high school students. Using this information, they drafted multiple-choice (MC) and constructed-response (CR) items that incorporated the targeted ideas and used published scientific data. Designing the assessments in this way reduced the likelihood of the assessments being biased and overly aligned to the curriculum. The items were piloted with middle and high school students (Homburger *et al.*, 2019).

The project's evaluators analyzed the alignment between the assessment items and the curriculum's NGSS learning goals. The analysis indicated a need for additional assessment items targeting the practice of argumentation and the topic of speciation. The assessment team developed an additional 20 items and pilot-tested them with 6191 high school students in 44 teachers' classrooms.

The GSLC curriculum development team did not see the items during the assessment development process or during any phase of curriculum testing. This helped ensure that the new unit was not unfairly aligned with the assessment and, thus, that the assessment was fair for both treatment and comparison conditions.

In the final version of the assessment, items were distributed across four online test forms, which were used for both pretesting and posttesting. Students were randomly assigned one form as a pretest and randomly assigned one of the three remaining forms as a posttest. Each form contained 26 MC items and two CR items, which assessed students' ability to write an argument. All test forms were designed to have the same average test difficulty, and each test had approximately the same number of items on each topic (see Appendix D in the Supplemental Material for the number of MC and CR items by topic). Student pre/posttests contained one common ancestry CR item and one natural selection CR item. The common ancestry CR item was the same on students' pre/posttests, while the natural selection CR item was different and asked students about slightly different scenarios in which natural selection occurred. Students could earn up to 12 points for each CR item. The 12-point rubrics evaluated the specificity, sophistication, and clarity of students' claims, evidence, and reasoning. The reliability of a rubric was evaluated a priori by three scorers independently scoring a subset of student responses. Scorers met to revise and clarify the rubric elements until they were able to obtain high interrater reliability (>0.80) for each rubric element. Additional information on the assessments, including the items, test forms, and summary results for each item for the treatment and comparison groups, can be found at http://assessment.aaas.org/topics/3.

Data Analysis. Only items that measured NGSS disciplinary core ideas for biological evolution (HS-LS4.A, B, C, D) were included in the assessment data that were analyzed for this study. The associated science practices of argumentation and analyzing data also were included in this analysis. No items that measured NGSS ideas for heredity/genetics (LS1.A, LS3.A, and LS3.B) were used in the data analysis, as students in the comparison group did not receive instruction on those topics. Comparison between the treatment and comparison groups involved a combination of Rasch modeling and hierarchical linear modeling (HLM).

Rasch Modeling. Rasch modeling was used first to evaluate the reliability of the assessments and whether or not they were suitable for making inferences about students. Second, Rasch modeling was used to obtain student ability measures for each student. In the Rasch model, the higher the student ability measure, the more likely a student is to answer a question correctly. Because these ability measures are obtained from a set of items, they give a measure of students' ability, or knowledge, associated with a set of items (Boone, 2016). Rasch analysis was conducted using the software package WINSTEPS (Linacre, 2019).

To evaluate the reliability of the items on the pre/posttest, we fit the pretest, posttest, and combined pre/posttest data to three separate Rasch models. After fitting each data set, we checked each item's fit to the model and the overall reliability of the model.

To obtain student ability measures on the pretest and posttest for each student, we combined the pre/posttest data by "stacking" them. Stacking the data allows for pre/posttest student measures to be measured on the same scale. This allows for a direct comparison of student measures on the pre/posttest, even though students completed different items on the tests.

Rasch analysis of the CR data was done by treating each element of the CR rubric as a dichotomous item. Initially, all rubric elements were included in the model. However, we found that some rubric elements had a poor fit. Specifically, rubric elements that gave students points for explicitly stating or defining evolution ideas did not fit to the model. Students typically did not describe natural selection or common ancestry and then apply those ideas. Because these rubric elements did not fit to the model, they were removed. The CR data were therefore modeled using only rubric elements that related to stating claims, citing evidence, and applying reasoning that links students' evidence to their claims.

Hierarchal Linear Modeling. Next, we evaluated the effect of the curriculum on students' posttest Rasch measures compared with the comparison group. Because student and classroom performance varies and depends on many factors, we fit student measures from the MC and CR items to an HLM model. This approach allowed us to determine the significance of the treatment on students' gains while accounting for the variation in student and classroom performance. We used a random intercept HLM model with a fixed treatment effect and pretest covariate. Students' pretest performance and whether their class was in the treatment condition were treated as fixed effects, while the unexplained variation in performance between different classes was treated as a second-level random effect. MC and CR measures were each fit to the model separately to obtain individual treatment effects for these two measures. The mathematical representation of the model is shown below.

Level 1—Student

$$\text{Posttest}_{ij} = \beta_{0j} + \beta_{1j} {*}\text{Pretest}_{ij} + r_{ij}$$

Level 2—Class

$$\beta_{0j} = \gamma_{00} + \gamma_{01} {*}\text{TREAT}_{ij} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

Teacher Implementation Data

We used two data sources—daily teacher logs and an end-of-unit survey—to understand and compare the critical features of the new and comparison units. Teachers in both conditions received slightly different daily teacher logs (see Appendices E and F in the Supplemental Material). However, parallel questions allowed us to compare the critical features of the different curricula used. Both versions of the logs asked which lesson(s) were taught and the level of emphasis (major, minor, touched on slightly, or none) placed on the NGSS practices and crosscutting concepts that day. The comparison teachers' logs also included fixed-choice questions about the content they taught and the materials they used in their lessons (i.e., textbook, readings outside the textbook, videos, online interactives, and other online materials). In addition, both logs included checkboxes with 15 options for the types of activities students participated in that day (e.g., making arguments using evidence, using digital technology, having small-group discussions about lesson content).

Treatment and comparison teachers received slightly different end-of-unit surveys. Both surveys asked open-ended questions about teachers' attitudes toward the units they taught (e.g., favorite and least favorite parts). The comparison condition survey included a question about whether students directly engaged with phenomena (e.g., through laboratory experiences). The treatment condition survey asked teachers about their intentions to use the unit in future years and how the unit differs from other units they have used to teach similar content.

We reviewed the teacher logs and end-of-unit surveys for evidence of content and activity implementation that would differentiate the treatment from comparison lessons. We defined "program" differentiation as the presence or absence of "critical features that distinguish the program from the comparison condition" during implementation (O'Donnell, 2008, p. 34). Content was one key factor, as we asked the comparison teachers to cover only evolution topics in their instruction, and not molecular genetics. We also hypothesized that the instructional activities would vary across conditions. For example, the new unit scaffolds students through the science practice of argumentation and includes activities on bar and line graphs. We asked: Would the comparison lessons provide opportunities to engage in the practice of argumentation or work with data? To answer these questions, we conducted an exploratory analysis of the teacher logs.

Log Selection. The treatment teachers submitted 424 logs (range 3–54 logs) and the comparison teachers submitted 315 logs (range 2–33 logs; see Appendix G in the Supplemental Material). These logs serve as a proxy for the amount of time teachers spent on their curricula. Because the logs varied by lesson content and in number per teacher, we reduced the sample to conduct a fair comparison. Similar to our approach with the student assessments, we tried to keep the content consistent across conditions. To do this, we restricted the treatment lessons to only those that addressed evolution topics (common ancestry, natural selection, and speciation) and excluded logs that focused only on molecular genetics topics. Second, to focus on the instruction, we eliminated logs in both conditions that 1) did not include any teaching (e.g., snow day, special event), 2) included only unit review, and 3) included only pre/posttesting. After restricting the sample, the number of logs per teacher still varied substantially within study conditions. However, log frequency did not vary significantly between groups, $t(36) = 1.54$, $p = 0.133$, suggesting that, as a group, the treatment and comparison teachers spent equal amounts of time on their evolution lessons.

Finally, we standardized the number of logs per teacher after establishing a minimum expectation for sampling. This ensured that the sample would not be weighted in favor of teachers with the largest number of lessons. We set the minimum number of logs per teacher at nine. While we asked teachers to complete one log for each lesson they taught, we expected that, at a minimum, teachers would complete two or three logs per week for at least 3 weeks. Some teachers taught in block schedules and did not meet with their students every day, making our estimate that much more realistic.

We eliminated eight teachers (four from each condition) with fewer than nine logs. Then we randomly selected nine logs from each remaining teacher using an online program

(Urbaniak and Plous, 2013). Our total lesson log sample was 270 (9 logs × 15 teachers per condition × 2 conditions).

Log Analysis. We conducted chi-square analysis to compare teachers' responses about the four science practices and crosscutting concepts emphasized in the treatment condition and the checkbox list of classroom activities (the 15 options for the types of activities students participated in that day). These analyses helped us to understand the frequency of each type of practice, crosscutting concept, and instructional activity in the treatment and comparison conditions. We made Bonferroni corrections to account for multiple comparisons, setting the p value 0.013 for the science practices and crosscutting concepts (0.05 divided by four comparisons), and 0.003 for the classroom activities (0.05 divided by 15 comparisons). We obtained Cramér's v and phi coefficients as measures of effect size. These coefficients are measures of association for 2 × 3 and 2 × 2 contingency tables, respectively. Their values range from −1 to 1, with absolute values of 0.2, 0.5, and 0.8, which generally indicate small, medium, and large effects (Ferguson, 2009). We chose to compare the effect sizes within the sample, identifying relatively larger and smaller effects among the surveyed items.

End-of-Unit Survey Analysis. We used three open-ended questions from the end-of-unit survey as a second method of differentiating the treatment and comparison units. Unlike the log analysis, these responses provided us with qualitative data. We asked teachers:

- How would you describe this unit to other biology teachers?
- What did you like the most about the unit you taught?
- What did you like the least about the unit you taught?

We assumed that teachers would mention the most salient features of the instructional materials. By analyzing their responses, we hoped to gain insights into what teachers might consider to be the distinguishing features of the treatment and comparison curricula as a whole.

We analyzed the three questions as a set through a combination of a priori codes and ones that emerged from the data. We generated codes for the content, practices, and activities, along with the positive and negative attributes of the curriculum. One primary rater (KMB) coded all responses (n = 36 total, 18 per condition). A second rater (DD-E) coded a random 25% sample to establish interrater reliability, defined in this study as percentage of exact agreement (Stemler and Tsai, 2008). Rater agreement for all three questions exceeded the 80% minimum reliability threshold.

RESULTS

Student Assessments

Figure 1 shows the MC and CR pre/posttest percentage correct for the treatment and comparison groups. There was a significant difference in percentage correct on MC items for treatment and comparison groups on the pretest, $t(2267) = -3.5$, $p < 0.000$, and posttest, $t(2267) = 8.6$, $p < 0.000$. Similar differences were observed for the pretest, $t(2267) = -3.4$, $p < 0.000$, and posttest, $t(2267) = 5.5$, $p < 0.000$, on the CR items.

These findings indicate that students in the treatment and comparison groups differed in their understanding of evolution concepts both before and after instruction. To better account for these differences and to determine whether the curriculum had an impact on student performance, we first used Rasch modeling to obtain scaled student measures for each student based on performance on the MC and CR items. We then used HLM modeling to control for variability in the student's posttest performance, including controlling for their pretest performance.

Rasch Modeling

All items had infit between 0.7 and 1.3, indicating a good fit to the Rasch models (Boone, 2016). In both models, a few items had outfit values larger than 1.3, likely due to some low-performing students correctly guessing the item. After examining the items with larger outfit values, we did not see any obvious reason for their slight overfitting. As the items had good infit statistics and including items with outfits values higher than 1.3 but lower than 2 was not degrading to the overall measurement

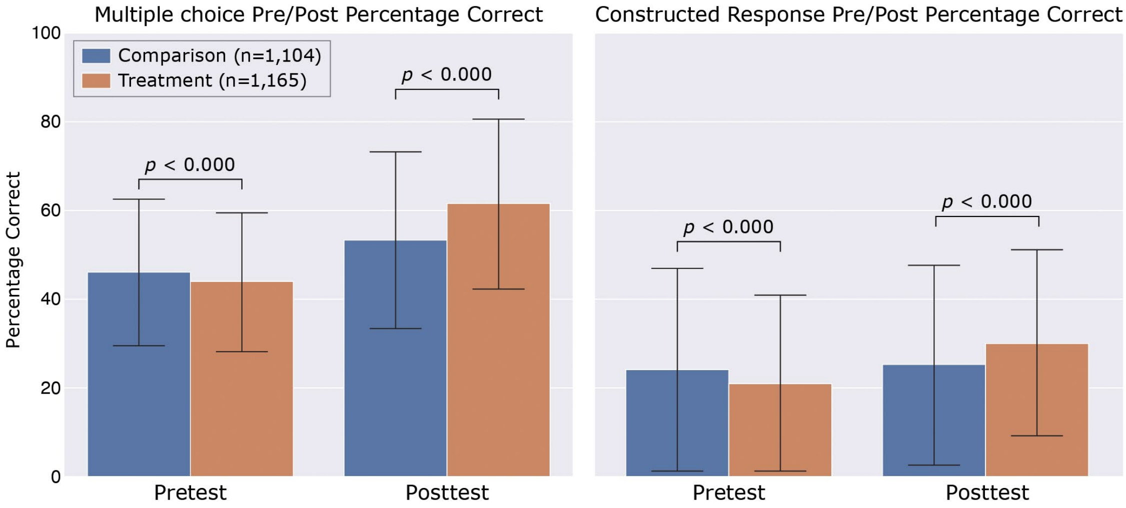

FIGURE 1. Percentage correct on the MC and CR tests for the treatment and comparison groups. Error bars indicate ±1 SD.

TABLE 1. Results from HLM model fit to MC measures

Fixed effects	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p* value
Intercept, γ_{00}	0.238	0.109	2.19	2230	0.029
Treatment, γ_{01}	0.355	0.154	2.31	36	0.027
Pretest, γ_{10}	0.562	0.024	23.2	2230	<0.001
Random effects	**Standard deviation**	**Variance**			
Intercept, u_0	0.447	0.200			
level 1, r	1.000	1.000			

(Linacre, 2002), we kept the items as part of the assessment. However, the person separations and reliabilities were relatively low, indicating the items may not be sensitive enough to distinguish between multiple strata of performers.

Hierarchical Linear Modeling

Tables 1 and 2 summarize the fit of the MC and CR student measures to the HLM model. Each fit of the student measures to the model indicated there was a significant difference ($p < 0.05$) in the posttest measures between the treatment and comparison groups. This indicates that students in the treatment group performed statically better on MC items, $t(2231) = 2.31$, $p = 0.027$, and on the CR items, $t(2231) = 2.23$, $p = 0.032$. We estimated a treatment effect size for each set of measures by dividing the treatment coefficient by the pooled SD in the posttest measures. This gave a Cohen's *d* effect size of 0.28 ± 0.12 for the MC posttest measures and 0.29 ± 0.13 for the CR posttest measures. Overall, these results indicate that students in the treatment group improved more in their performance than students in the comparison group on both the content knowledge–focused MC items and the argumentation-focused CR items.

Teacher Implementation

We report two sources of information about program differentiation: 1) teachers' activity records from their lesson logs and 2) teachers' descriptions of their units from their end-of-unit surveys.

Log Analysis. Recall that, in the lesson logs, teachers noted the types of instructional activities they did each day, selecting from a list of 15 choices (e.g., use computers or online lessons, engage in small-group discussion, analyze and interpret data). We hypothesized that, consistent with the curricular design principles, teachers in the treatment condition would give students more opportunities to engage in the science practices of engaging in argumentation from evidence and analyzing and interpreting data. We also expected that the treatment teachers would use educational technology more frequently by virtue of their access to the unit's videos, interactives, and simulations.

As expected, chi-square analysis revealed that teachers who used the new curriculum had their students use computers or online lessons and view videos more often than the teachers in the comparison condition (Table 3). Analyses of these activities generated statistically significant differences with small effect sizes. In contrast, differences between conditions in the frequency of argumentation and data analysis were not statistically significant. We found the largest difference between the two conditions in the amount of large- and small-group discussions. These occurred in ~90% of the treatment lessons and 60% of the comparison lessons.

Teachers also rated the level of emphasis they placed on the two science practices and two crosscutting concepts featured in the treatment materials. While the frequency of opportunities to engage in argumentation did not vary by study condition, the relative emphasis on that practice did vary. A "great deal" of emphasis on argumentation was included in 61.4% of the treatment lessons compared with 38.5% of the comparison lessons. This small effect was statistically significant (Table 4).

General Unit Descriptions. We coded the content, practices, activities, and resources teachers mentioned in their responses as a way to identify what they may have considered to be the key components of the units (Table 5).

Across the three questions (i.e., unit description, like most and like least about the unit they taught), we found differences in the kinds of details teachers in the treatment and comparison conditions shared about their units. Nearly all (94.4%) of the comparison teachers described the evolution topics they taught, including natural selection, fossil evidence for evolution, and Darwin's theories. Only about three-fifths of the treatment teachers wrote about the content, representing a statistically significant, moderately-sized difference between conditions; $\chi^2(1) = 5.79$, $p = 0.016$, effect size (φ) = 0.401. Conversely, treatment teachers were twice as likely as the comparison teachers to report the NGSS science practices in their unit

TABLE 2. Results from HLM model fit to CR measures

Fixed effects	Coefficient	Standard error	*t* ratio	Approx. *df*	*p* value
Intercept, γ_{00}	–2.078	0.198	–10.5	2230	<0.001
Treatment, γ_{01}	0.625	0.280	2.23	36	0.032
Pretest, γ_{10}	0.511	0.019	26.5	2230	<0.001
Random effects	**Standard deviation**	**Variance**			
Intercept, u_0	0.823	0.677			
level 1, r	1.658	2.749			

TABLE 3. Frequency of instructional activities by study condition

Activity	% Yes, treatment (*n* = 135 lessons)	% Yes, comparison (*n* = 135 lessons)	χ^2 (1)	*p*	Effect size (φ)
Have large-group discussions about today's lesson topic	90.2	58.2	35.32	0.000*	0.364
Have small-group discussions about today's lesson topic	87.9	62.2	23.37	0.000*	0.296
Work with bar graphs	40.8	18.8	15.21	0.000*	0.241
Use computers or online lessons	49.6	30.1	10.52	0.001*	0.200
View a video or film	66.7	47.7	10.10	0.001	0.194
Work with line graphs	33.1	17.3	8.71	0.003	0.182
Read a handout or assigned text	83.3	68.9	7.64	0.006	0.169
Analyze or interpret data	82.6	68.2	7.37	0.007	0.167
Make arguments using evidence	72.7	60.4	4.50	0.034	0.130
Collect data	38.6	28.8	2.87	0.091	0.104
Evaluate their own arguments	53.0	43.0	2.71	0.100	0.101
Evaluate their peers' arguments	36.4	29.9	1.29	0.257	0.070
Do a hands-on activity (e.g., in a lab or by observing or working with organisms)	40.3	39.8	0.006	0.939	0.005
Develop or interpret models (e.g., a picture or animation of a food chain, or a drawing of an atom)	46.9	48.1	0.04	0.846	–0.012
Do an assessment (e.g., test, quiz, exit ticket)	26.9	41.5	6.23	0.013	–0.153

[a]*p* value is equal to or smaller than the Bonferroni-adjusted alpha = 0.003.

(66.7% vs. 27.8%), with argumentation from evidence being the most frequently mentioned.

Consistent with the lesson logs, we found no statistically significant difference between frequency of hands-on activities across conditions; $\chi^2(1) = 1.78$, $p = 0.182$. References to class discussions, a prominent feature of the individual lesson logs, appeared only three times in the end-of-implementation responses, and only in the treatment teacher responses (16.7% of the study group).

Many of the positive and negative curricular attributes that teachers reported focused on the integration of molecular genetics and evolution, or the lack thereof. (As a reminder, the comparison teachers were restricted from teaching a molecular genetics unit during the school year until after the posttest.) Eight of the treatment teachers (44.4%) wrote about the unit's coherence. Five of these explicitly described the advantages of teaching evolution with a grounding in molecular genetics. As two teachers noted,

> "The Evolution unit is a really interactive, novel approach to teaching evolution through genetics. Genetics play such an important role in evolution and this unit really showcases that fact so well."

> "I would describe this unit as a NGSS-aligned curriculum that focuses on big ideas and skills in science through real life data analysis. It aligns content that you wouldn't necessarily teach under the umbrella of evolution, like transcription and translation, but does it in a seamless way that allows students to fully understand the mechanics of evolution on a micro and macro level."

TABLE 4. Comparison of level of emphasis on science practices and crosscutting concepts by study condition

Lesson emphasis	*n*	χ^2 (3)	*p*	Effect size (Cramér's *v*)
Engaging in argument from evidence	267	14.34	0.002*	0.232
Analyzing and interpreting data	265	7.72	0.052	0.171
Patterns	263	6.43	0.092	0.156
Cause and effect	263	3.88	0.275	0.121

**p* value is equal to or smaller than the Bonferroni-adjusted alpha of 0.013.

Seven of the comparison teachers (38.9%) bemoaned students' difficulty understanding evolutionary concepts without a background in molecular genetics. As one teacher wrote,

> "I would not recommend teaching evolution before having taught about DNA or heredity. There were times the students got caught up on some of the terminology and did not always have a deeper understanding. I could tell they struggled with the idea of genetic variation because they did not yet know about genetic recombination, alleles, etc. Although I think they did okay, but not with as deep of an understanding as I would like."

The treatment teachers' survey also asked whether they planned to teach the unit or parts of it the following year, and, if so, in what way:

- All (100%) of the teachers indicated that they will use the unit in some way;
- 62% planned to teach it in sequence with the addition of other curricular materials, including labs such as DNA extraction, fossils, and descriptions of Darwin's voyage;
- 28% planned to teach select parts of the unit;
- 5% planned to teach the entire unit but in a different sequence; and
- 5% planned to teach the entire unit in sequence but divided into parts during the year.

Collectively, the lesson log and end-of-implementation survey results suggest that the new evolution unit varies from the comparison units in several ways. While the frequency of

TABLE 5. Frequency of curricular features by study condition

Activity	% Yes, treatment (n = 18 responses)	% Yes, comparison (n = 18 responses)	χ^2 (1)	p	Effect size (φ)
Description of evolution content	58.8	94.4	5.79	0.016	–0.401
Description of any NGSS science practices	66.7	27.8	5.46	0.019	0.389
Engagement in argumentation from evidence	50.0	0.0	12.00	0.001	0.577
Use of hands-on activities	61.1	38.9	1.78	0.182*	0.222

[a]p value is equal to or smaller than the Bonferroni-adjusted alpha = 0.0125.

opportunities to engage in argumentation from evidence was similar across conditions, teachers in the treatment condition placed a greater emphasis on this practice. In addition, the new unit enabled students to make sense of phenomena through discussion (the teacher guides provide discussion questions related to the phenomena explored in each lesson). These distinctions may be associated with the differential gains in student performance.

DISCUSSION

The findings from this study with 2269 students and their 38 teachers provide support for our hypothesis that a cohesive evolution unit that uses the principles of NGSS and integrates across biological organizational levels fosters better learning outcomes and teacher experiences than teachers' typical evolution curricula. In addition, the results offer insight into evolution teaching and learning and selection of curricular materials for both secondary and postsecondary educators and researchers.

Comparison of the Conditions

Multiple, interrelated curricular components likely contributed to the observed results. The GSLC developed a three-dimensional curricular unit that intertwined evolution content with the science practices of argumentation and data interpretation and the crosscutting concepts of patterns and cause and effect. While we randomly assigned teachers to include (treatment) or exclude (comparison) molecular genetics content, we cannot attribute score differences to that content alone. The implementation surveys revealed some additional key differences in instructional activities, suggesting that *how* teachers taught was as relevant as *what* they taught. We consider three of the most distinct, measured differences between the two conditions, while acknowledging that the complex interaction of these factors likely contributed to student performance.

The implementation data, summarized in Table 6, showed that students in the treatment condition engaged in science practices and whole- and small-group discussions more often than students in the comparison condition. Treatment students used technology-based lessons more frequently, while comparison students completed more assessments. Both conditions placed approximately equal emphasis on crosscutting concepts, use of hands-on activities, evaluating arguments, and collecting data. These similarities and differences and their potential impact on student learning are discussed in the next four subsections.

Integrating Molecular Genetics. Our study findings support those from other research, which has shown increased student understanding of evolution when molecular genetics was part of an evolution curriculum compared with when it was not (e.g., Jördens *et al.*, 2016). The qualitative data from our end-of-implementation teacher survey complement our quantitative results. Many of the treatment teachers reported on the advantages of grounding evolution in molecular genetics for understanding the mechanics of evolution. The comparison teachers, in turn, described the difficulty of teaching evolution without a background in molecular genetics.

Opportunities for Science Practices. We found that the treatment lessons emphasized argumentation from evidence more than the comparison lessons. Treatment students also worked with bar and line graphs (related to analyzing and interpreting

TABLE 6. Key similarities and differences between conditions

Element	Treatment condition	Comparison condition
Time spent on evolution topics	~ 5 weeks	3–6 weeks
Disciplinary core ideas for evolution	Yes	Yes
Disciplinary core ideas for molecular genetics	Yes	No
Classical genetics	No	No
Science practice: analyzing and interpreting data through working with line and bar graphs[a]	More often	Less often
Science practice: evidence-based argumentation[a]	More often	Less often
Technology-based lessons (computers, online lessons, videos)[a]	More often	Less often
Topic-relevant discussions (whole and small group)[a]	More often	Less often
Crosscutting concepts: patterns, and cause and effect	Equal	Equal
Use of hands-on activities	Equal	Equal
Evaluate own or peers' arguments	Equal	Equal
Collecting data	Equal	Equal
Assessments (tests, quizzes)	Less often	More often

[a]Statistically significant differences.

data) more frequently. We speculate that increased opportunities to carry out these practices contributed to student learning.

No significant differences were found in providing students opportunities to evaluate their own arguments, evaluate one another's arguments, or collect data. Both conditions used these practices, which are important in scientific sense-making. This suggests that the processes of collecting data and evaluating one's own or another's arguments may not advance evolution understanding as much as arguing from evidence and analyzing and interpreting data.

The significant differences between the conditions are revealing and potentially tell a story of practices that may benefit student learning of evolution. This supports previous research that has shown students' understanding of evolution improves when they use curricula that incorporate argumentation (e.g., Bell and Linn, 2000; Zohar and Nemet, 2002; Asterhan and Schwarz, 2007) and provide opportunities to analyze and interpret data (e.g., Bray Speth *et al*., 2009; Beardsley *et al*., 2011). Based on these findings, the GSLC unit developers selected argumentation and working with data as the unit's target practices. Each module includes multiple opportunities for students to analyze and interpret skill level–appropriate data about phenomena from published scientific research. Students learn how to construct evidence-based arguments based on these phenomena, using a scaffolded approach that builds across all five modules. For example, in the first module, to establish that all living things share the same basic biochemistry, students evaluate arguments about bioengineering approaches, insulin produced by yeast, and the use of green fluorescent protein from jellyfish in other animal models for research. In the second module, students complete arguments about the ancestry of cetaceans by selecting relevant evidence from four lines: anatomy, fossils, embryos, and DNA. By the fifth module, students independently write an argument about a potential speciation event in flies using evidence they examine from observational and genetic studies.

Opportunities for Discussion and Assessment. Another significant difference between the conditions was around class discussion. Treatment students engaged in large- and small-group discussions in nearly every lesson (90% and 88%, respectively), which was significantly more often than the comparison students (58% and 62%, respectively).

In the treatment unit, each module is organized around overarching and guiding questions with the idea that teachers will use them as discussion prompts to check for understanding as they progress through the unit. The overarching questions connect the core ideas explored in each module to molecular genetics concepts, while the guiding questions in each module support the overarching question. For example, the overarching question for the module that explores the shared biochemistry of living things asks: "What shapes the characteristics of all living things?" The supporting questions that guide the lesson sequence are: "Why can living things decode the information in each other's genes?" and "If organisms build proteins the same way, do they build the same proteins?" The module's lessons guide students in discovering that DNA and the proteins it encodes are responsible for the characteristics of all living things and that there are protein-level similarities among even vastly different organisms.

The necessity of breaking up large data analysis tasks into smaller chunks naturally builds in small-group discussion (doing the analysis task) and sharing in large-group settings. Also, in some cases, whole-group discussion questions are suggested as a way for teachers to do a quick check for understanding before moving on to topics that build on the content that was just learned. The discussion prompts require students to use molecular genetics to make sense of evolution. We speculate that the GSLC unit may have provided opportunities for student-centered discussion and sense-making of evolutionary phenomena in this way.

The results showed no statistically significant association between study condition and frequency of assessment practices such as tests, quizzes, or exit tickets. It is possible that the treatment teachers may have interpreted the question as assessments *in addition* to those already in the curriculum (the curriculum has formative assessments embedded in each module and an end-of-unit summative assessment). Further, as described earlier, the treatment condition showed significantly more large- and small-group discussions than the comparison condition, which teachers may have used as formative assessments to help uncover student thinking.

Opportunities for Technology Use. Finally, the implementation data revealed that students in the treatment condition used computers or online lessons and viewed videos significantly more often than students in the comparison condition. In the comparison condition, teachers most frequently used videos and online interactives from HHMI Biointeractive (HHMI, nd) and PBS (PBS, nd).

In the treatment unit, videos are used to introduce phenomena and engage students' interest in them. Animations depict dynamic or molecular-level processes. "Click-through" interactive lessons provide summaries of key concepts or address misconceptions, allowing for discussion pauses, checks for understanding, or the opportunity to review.

Our finding of increased student learning with the treatment unit aligns with the literature on computer technologies promoting student-centered learning (Ertmer *et al*., 2012; Hechter and Vermette, 2014). These technologies allow students to observe or interact with complex and dynamic biological processes. In addition, video can be a rich and powerful medium, because it can present information in an attractive manner (Wieling and Hofman, 2010). We speculate that production of videos specifically designed for the treatment unit, as well as the opportunities to observe dynamic and molecular-level processes through interactive multimedia, slideshows, and videos likely benefited student learning in the treatment condition. Further, technology-based lessons are sometimes more time effective than paper-based lessons (e.g, Miller *et al.*, 2018). The GSLC unit's technology pieces may have helped teachers cover more information in each class period. We speculate that the delivery of information was more efficient than in the comparison lessons, which may have contributed to increased learning gains in the treatment group.

The Effect of Curricular Coherence

We speculate that the overall coherence of the GSLC's unit may also have supported student learning. Coherent instructional materials take what would otherwise be disconnected pieces of information and unite them (NGSS Lead States, 2013). A coherent unit builds knowledge and skills over lessons, units, and time (Fortus and Krajcik, 2012) and aligns target science ideas and varying depths of those ideas (Schmidt *et al.*, 2005).

In the "Evolution: DNA and the Unity of Life" unit, the developers used the principles of curricular coherence to achieve their goal of effectively bringing molecular genetics and evolution together. They carefully selected their methods of integrating molecular genetics and evolution topics based on learning progression literature (see Homburger *et al.*, 2019). Key molecular genetics concepts and processes are repeated or referred to in lessons throughout the unit at relevant points. In this way, the unit establishes a common "molecular genetics language" through which to convey evolution concepts. As described in the introductory section of Homburger *et al.* (2019), the GSLC team's science content experts, along with scientists and teachers, determined which molecular concepts reflected the literature on student sense-making while learning evolution.

The developers used the approach advocated by Fortus and Krajcik (2012) around building curricula that progress from simpler to more complex levels of understanding and specifying how students should use these understandings. Lessons in the final module of the unit ("Speciation") bring together the key concepts from the previous four modules while tasking students with connecting those concepts to explain the divergence in characteristics of living things. Key module-level learning goals inform embedded formative assessments in each module. Further, the developers embedded these assessments strategically to measure student learning while teachers still had time to use the feedback to inform instruction. This speaks to the frequent use of discussion in the treatment condition. Finally, it is likely that the development team's multiple rounds of classroom testing and iterative revision processes streamlined the treatment unit and contributed to its coherence (Homburger *et al.*, 2019). Perhaps as a result, nearly half of the treatment teachers in our study mentioned the coherence of the content in the end-of-implementation survey. While we did not directly examine the coherence of the comparison conditions, the treatment teachers' comments suggest that they found this to be a distinctive feature.

Study Limitations

As with most RCT designs in educational settings, this study has several limitations worth highlighting. First, the research design and the treatment unit do not allow us to definitively separate the effects of integrating molecular genetics on student learning outcomes. To understand this effect, our design would have needed to include a comparison condition in which the new unit was taught without the integrated molecular genetics components. However, this was not possible to do with a unit in which these concepts are so tightly interwoven. Another approach to testing this would have been to create subtests in order to obtain separate measures for each topic. We could have used these measures in regression or path analysis models to examine how the different measures correlate with one another. However, we determined that these measures have poor reliability and are not distinct. These regression models would result in a large error in the fit statistics. Another possibility to control for molecular genetics more tightly would have been to provide comparison teachers with a unit for everyone to use that did not include molecular genetics. However, this approach would have involved developing an additional unit, which was outside the scope of funding for this project.

Second, identifying a fair comparison or comparison group is difficult and often problematic (Drits-Esser *et al.*, 2014). As such, we had to consider what constituted a fair comparison group that was practical for teachers. We determined that we could come close by aligning the treatment unit's NGSS connections to the requirement for the comparison connections. Our definition of an NGSS comparison unit was limited to "using the same evolution disciplinary core ideas as the treatment unit." We based our guideline for establishing an NGSS comparison unit on teachers' answers to questions on the application about their current use of NGSS.

Our review of the literature suggested that most RCT studies that investigate the efficacy of NGSS-oriented curricula used teachers' standard curricula (also called business as usual) as comparison conditions (e.g., Hand *et al.*, 2018; Taylor *et al.*, 2015; Batiza *et al.*, 2016; Llosa *et al.*, 2016; Rachmawati *et al.*, 2019; Kim *et al.*, 2020; Schiefer *et al.*, 2020). This type of comparison condition is common, because it can provide evidence that the innovative treatment was superior to what is currently considered standard practice. While identifying a fair comparison group is difficult and no solution is perfect, our decision to compare the GSLC's innovative NGSS-friendly curriculum with existing NGSS comparison curricula is an accepted method of conducting an RCT study.

A third limitation revolves around the problematic nature of measuring implementation. We had to develop a reasonable, cost-effective research design that would measure implementation in classrooms across the United States. Teacher self-report includes bias and cannot capture everything that may be important to note. However, for our study, this method of data collection was the most practical, as we were unable to conduct classroom observations in multiple states.

Finally, additional implementation factors may have influenced the results. First, the treatment group used one set of curricular materials, while the comparison group used many different types of textbooks and other materials. Thus, there was more variation in the comparison group, which likely led to an underestimation of effect size (Kraft, 2020). Second, there may have been a practice effect in favor of the comparison group. The treatment teachers were using a new curriculum, while the comparison teachers were using units they had previously implemented. Ideally, we would have conducted the study a year after treatment teachers first implemented the new unit to at least partially stave off the practice effect. Third, there may have been a Hawthorne effect, or an effect of being monitored or observed on outcomes. The teachers knew they were participating in a research study *and* were self-reporting their practices, which may have caused them to do the "best" implementation of the curricular materials that they could. This could favor both conditions equally.

Future Research

Additional research is needed on molecular genetics and evolution in curricula using different research designs, methods, and comparison groups. While the treatment unit provides one model for successful integration, there are other methods to study. This growing field still has much to learn.

The limitations in this study leave the door open for interested investigators to conduct further research into many aspects of the "Evolution: DNA and the Unity of Life" unit. For example, further investigation into the implementation results would add to our understanding of the unit's effects. This could include understanding the specific differences in the types of conversations that occurred during large- and small-group discussions. What did students discuss, and what science practices did those discussions reflect (argumentation, explanation, computational thinking, etc.)? How—if at all—do students incorporate information from prior modules and lessons into their discussions? What evidence do these discussions provide about the progression of students' evolutionary reasoning? What is the instructional experience of students who are members of groups that are underrepresented in evolutionary science (Mead *et al*, 2015; Barnes *et al*., 2020)?

In our study, teachers implemented the treatment unit under closely prescribed conditions. We asked them to teach the curriculum strictly in sequence, and not add, delete, or substitute activities. When asked how they would teach the curriculum in the future, however, some respondents indicated that they would supplement it with other materials or somewhat alter the sequence. It is expected and even desired for teachers to adjust their instruction to accommodate the prior knowledge and experiences of their students. However, the extent to which these adaptations maintain the activities' core cognitive demands and facilitate the integration of molecular genetics and evolution is unknown (Tekkumru-Kisa *et al.*, 2020). Future research could investigate these variations to determine their effect on students' evolution knowledge and reasoning. It also would be helpful to identify professional development that can best support teachers in integrating molecular genetics and evolution in their instruction.

CONCLUSIONS

The results from this national randomized controlled trial study revealed that students exposed to the "Evolution: DNA and the Unity of Life" curriculum outperformed students who learned with their teachers' regular evolution curricula. This was true for both measures of evolution content and for argumentation from evidence. Our findings suggest that this outcome could be due to a combination of cognitive advantages from using a curricular unit that emphasizes and scaffolds constructing arguments from evidence, involves analyzing and interpreting skill level–appropriate data about phenomena, provides frequent opportunities to engage in large- and small-group discussions, and incorporates technology. Further, the unit was developed using principles of curricular coherence to integrate molecular genetics and evolution throughout.

The comparison units provided fewer opportunities for students to engage in these NGSS science practices, to participate in large- and small-group discussions, and to utilize technology. While the tightly integrated nature of the unit prevented us from teasing out the effects of each of these curricular components on student learning, our findings lend themselves to opportunities for further research into the ways in which these curricular elements support evolution learning.

ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation under grant numbers DRL-1418136 and DRL-1222869. 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. We are grateful to the many professionals who contributed to this project as well as the thousands of students who participated throughout development and testing of the unit and assessment items. A full list of acknowledgments, schools, and locations can be found at https://learn.genetics.utah.edu/content/evolution/credits. We are also grateful to the journal reviewers for their thoughtful, thorough, and gracious feedback throughout the entire revision process. Monitoring editor Tessa Andrews provided expert guidance. Her time and efforts contributed greatly to our work.

REFERENCES

Anderson, C. W., de los Santos, E. X., Bodbyl, S., Covitt, B. A., Edwards, K. D., Hancock, J. B., … & Welch, M. M. (2018). Designing educational systems to support enactment of the Next Generation Science Standards. *Journal of Research in Science Teaching*, *55*(7), 1026–1052. https://doi.org/10.1002/tea.21484

Asterhan, C. S., & Schwarz, B. B. (2007). The effects of monological and dialogical argumentation on concept learning in evolutionary theory. *Journal of Educational Psychology*, *99*(3), 626.

Barnes, M. E., Supriya, K., Dunlop, H. M., Hendrix, T. M., Sinatra, G. M., & Brownell, S. E. (2020). Relationships between the religious backgrounds and evolution acceptance of Black and Hispanic biology students. *CBE—Life Sciences Education*, *19*(4), 1–14.

Batiza, A., Luo, W., Zhang, B., Gruhl, M., Nelson, D., Hoelzer, M., … & Marcey, D., & Society for Research on Educational Effectiveness. (2016). Regular biology students learn like AP students with SUN. In *Society for Research on Educational Effectiveness (Spring conference), held March 2-5, 2016, Washington, DC*. Society for Research on Educational Effectiveness.

Beardsley, P. M., Stuhlsatz, M. A. M., Kruse, R. A., Eckstrand, I. A., Gordon, S. D., & Odenwald, W. F. (2011). Evolution and medicine: An inquiry-based high school curriculum supplement. *Evolution: Education and Outreach*, *4*(4), 603–612. https://doi.org/10.1007/s12052-011-0361-2

Bell, P., & Linn, M. C. (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal of Science Education*, *22*(8), 797–817. https://doi.org/10.1080/095006900412284

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education*, *15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Borgerding, L. A., Klein, V. A., Ghosh, R., & Eibel, A. (2015). Student teachers' approaches to teaching biological evolution. *Journal of Science Teacher Education*, *26*(4), 371–392.

Bray Speth, E. B., Long, T. M., Pennock, R. T., & Ebert-May, D. (2009). Using Avida-ED for teaching and learning about evolution in undergraduate introductory biology courses. *Evolution: Education and Outreach*, *2*(3), 415–428. https://doi.org/10.1007/s12052-009-0154-z

Drits-Esser, D., Bass, K., & Stark, L. A. (2014). Using small-scale randomized controlled trials to evaluate the efficacy of new curriculum materials. *CBE—Life Sciences Education*, *13*(4), 593–601. doi: 10.1187/cbe.13-08-0164

Ertmer, P. A., Ottenbreit-Leftwich, A. T., Sadik, O., Sendurur, E., & Sendurur, P. (2012). Teacher beliefs and technology integration practices: A critical relationship. *Computers & Education*, *59*(2), 423–435. https://doi.org/10.1016/j.compedu.2012.02.001

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538. https://doi.org/10.1037/a0015808

Ferrari, M., & Chi, T. H. (1998). The nature of naïve explanations of natural selection. *International Journal of Science Education*, *20*(10), 1231–1256.

Fortus, D., & Krajcik, J. (2012). Curriculum coherence and learning progressions. In Fraser, B., McRobbie, C., & Tobin, K. (Eds.), *Second international handbook of science education* (pp. 783–798). Dordrecht: Springer.

Genetic Science Learning Center (GSLC). (2018a). *Evolution: DNA and the unity of life (student site)*. Retrieved September 10, 2020, from https://learn.genetics.utah.edu/content/evolution.

Glaze, A. L., & Goldston, M. J. (2015). US science teaching and learning of evolution: A critical review of the literature 2000–2014. *Science Education*, *99*(3), 500–518.

GSLC. (2018b). *Evolution: DNA and the unity of life (teacher site)*. Retrieved September 10, 2020, from https://teach.genetics.utah.edu/content/evolution.

Gregory, T. R. (2009). Understanding natural selection: Essential concepts and common misconceptions. *Evolution*, *2*, 156–175.

Hand, B., Shelley, M. C., Laugerman, M., Fostvedt, L., & Therrien, W. (2018). Improving critical thinking growth for disadvantaged groups within elementary school science: A randomized controlled trial using the Science Writing Heuristic approach. *Science Education*, *102*(4), 693–710.

Hechter, R., & Vermette, L. A. (2014). Tech-savvy science education? Understanding teacher pedagogical practices for integrating technology in K–12 classrooms. *Journal of Computers in Mathematics and Science Teaching*, *33*(1), 27–47.

HHMI Biointeractive. (nd). *Home page*. Retrieved December 10, 2020, from www.biointeractive.org.

Homburger, S. A., Drits-Esser, D., Malone, M., Pompei, K., Breitenbach, K., Perkins, R. D., ... & Stark, L. A. (2019). Development and pilot testing of a three-dimensional, phenomenon-based unit that integrates evolution and heredity. *Evolution: Education and Outreach*, *12*, 13. https://doi.org/10.1186/s12052-019-0106-1

Jördens, J., Asshoff, R., Kullmann, H., & Hammann, M. (2016). Providing vertical coherence in explanations and promoting reasoning across levels of biological organization when teaching evolution. *International Journal of Science Education*, *38*(6), 960–992. https://doi.org/10.1080/09500693.2016.1174790

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*(2), 87–97. https://doi.org/10.1187/cbe.09-12-0088

Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., & Elmore, J. (2020). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology*, *113*(1), 3–26 (Supplemental). https://doi.org/10.1037/edu0000465.supp

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

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2019). *Winsteps Rasch measurement computer program.* Retrieved November 18, 2019, from www.winsteps.com

Llosa, L., Lee, O., Jiang, F., Haas, A., O'Connor, C., Van Booven, C. D., & Kieffer, M. J. (2016). Impact of a large-scale science intervention focused on English language learners. *American Educational Research Journal*, *53*(2), 395–424.

Mead, L. S., Clarke, J. B., Forcino, F., & Graves, J. L. (2015). Factors influencing minority student decisions to consider a career in evolutionary biology. *Evolution: Education and Outreach*, *8*(6). https://doi.org/10.1186/s12052-015-0034-7

Mead, R., Hejmadi, M., & Hurst, L. D. (2017). Teaching genetics prior to teaching evolution improves evolution understanding but not acceptance. *PLoS Biology*, *15*(5), e2002255. https://doi.org/10.1371/journal.pbio.2002255

Miller, T. A., Carver, J. S., & Roy, A. (2018). To go virtual or not to go virtual, that is the question: A comparative study of face-to-face versus virtual laboratories in a physical science course. *Journal of College Science Teaching*, *48*(2), 59–67.

National Research Council (NRC). (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: National Academies Press. https://doi.org/10.17226/13165

NRC. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.

NRC. (2015). *Guide to implementing the Next Generation Science Standards*. Washington, DC: National Academies Press. https://doi.org/10.17226/18802

Next Generation Science Standards Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: National Academies Press.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*(1), 33–84. https://doi.org/10.3102/0034654307313793

PBS. (nd). Retrieved December 10, 2020, from www.pbs.org/wgbh/evolution/change/family, http://www.pbs.org/wgbh/nova/labs/lab/evolution/research#/evo/buildatree/1, http://www.pbs.org/wgbh/nova/labs/about-evolution-lab/educator-guide/ .

Rachmawati, E., Prodjosantoso, A. K., & Wilujeng, I. (2019). Next Generation Science Standards in science learning to improve student's practice skill. *International Journal of Instruction*, *12*(1), 299–310.

Schiefer, J., Stark, L., Gaspard, H., Wille, E., Trautwein, U., & Golle, J. (2020). Scaling up an extracurricular science intervention for elementary school students: It works, and girls benefit more from it than boys. *Journal of Educational Psychology*, *113*(4), 784–807 (Supplemental). https://doi.org/10.1037/edu0000630.supp

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: An examination of US mathematics and science content standards from an international perspective. *Journal of Curriculum Studies*, *37*(5), 525–559. https://doi.org/10.1080/0022027042000294682

Stemler, S. E., & Tsai, J.& (2008). Best practices in interrater reliability three common approaches. In Osborne, J. (Ed.), *Best practices in quantitative methods* (pp. 29–49). Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781412995627.d5

Taylor, J. A., Getty, S. R., Kowalski, S. M., Wilson, C. D., Carlson, J., & Van Scotter, P. (2015). An efficacy trial of research-based curriculum materials with curriculum-based professional development. *American Educational Research Journal*, *52*(5), 984–1017.

Tekkumru-Kisa, M., Stein, M. K., & Doyle, W. (2020). Theory and research on tasks revisited: Task as a context for students' thinking in the era of ambitious reforms in mathematics and science. *Educational Researcher*, *49*(8), 606–617. https://doi.org/10.3102/0013189X20932480

Urbaniak, G. C., & Plous, S. (2013). *Research randomizer (Version 4.0).* Retrieved April 9, 2019, from www.randomizer.org

What Works Clearinghouse. (2017). *Standards handbook* (Version 4.0). Retrieved June 28, 2018, from https://eric.ed.gov/?id=ED577036

White, P. J. T., Heidemann, M. K., & Smith, J. J. (2013). A new integrative approach to evolution education. *BioScience*, *63*(7), 586–594. https://doi.org/10.1525/bio.2013.63.7.11

Wieling, M. B., & Hofman, W. H. A. (2010). The impact of online video lecture recordings and automated feedback on student performance. *Computers & Education*, *54*(4), 992–998. https://doi.org/10.1016/j.compedu.2009.10.002

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*(1), 35–62. https://doi.org/10.1002/tea.10008