Effectiveness of *Enhanced Units*

A REPORT OF A RANDOMIZED EXPERIMENT IN CALIFORNIA AND VIRGINIA

December 2019

Andrew P. Jaciw

Jenna Zacamy

Hannah D'Apice

Li Lin

Connie Kwong

Adam M. Schellinger

Empirical Education Inc.

Acknowledgements

We are grateful to the teachers and staff in participating schools in California and Virginia for their assistance and cooperation in conducting this research. The research was conducted under a subcontract with the Center for Applied Special Education Technology (CAST) through their partnership with SRI International on their 2014 Investing in Innovation (i3) Development grant (Award Number U411C140003). As the independent evaluator, Empirical Education Inc. was provided with independence in reporting the results.

About Empirical Education Inc.

Empirical Education Inc. is a Silicon Valley-based research company that provides tools and services to help K-12 school systems make evidence-based decisions about the effectiveness of their programs, policies, and personnel. The company brings its expertise in research, data analysis, engineering, and project management to customers that include the U.S. Department of Education, educational publishers, foundations, leading research organizations, and state and local education agencies.

Table of Contents

Reference this report:

Jaciw, A. P., Zacamy, J., D'Apice, H., Lin, L., Kwong, C., & Schellinger, A. M. (2019). *Effectiveness of Enhanced Units: A Report of a Randomized Experiment in California and Virginia*. (Empirical Education report number: Empirical_CAST_EU-7034-FR1-2019-O.3). San Mateo, CA: Empirical Education Inc.

Background

Empirical Education Inc. is the independent evaluator of SRI International's 2014 Investing in Innovation (i3) Development grant called Redesigning Secondary Courses to Improve Academic Outcomes for Adolescents with Disabilities and Other Underperforming Adolescents. The goal of the grant is to develop *Enhanced Units* that combine research-based content enhancement routines, collaboration strategy, and technology components for secondary U.S. History and biology classes. This report presents findings of a randomized control trial (RCT) during the 2017-18 school year. The RCT measured the impact of *Enhanced Units* on higher order content skills (as measured through unit tests) in high school biology and U.S. History classes in three districts in Virginia and California.

DESCRIPTION OF *ENHANCED UNITS*

SRI, the Center for Applied Special Education Technology (CAST), and their research and practitioner partners developed *Enhanced Units* (*EU*) with the goal of integrating research-based content enhancement routines (referred to in this report as *routines*) with technological enhancements to improve student content learning and higher order reasoning, especially for students with disabilities or other learning challenges. The routines used in the study are based on the Strategic Instruction Model (SIM) and were developed at the University of Kansas Center for Research on Learning. SIM interventions are based on the application of the principles of systematic, explicit, guided instruction, mastery of critical content, and the use of cognitive and metacognitive supports related to completing academic and social tasks that improve student learning. SIM lessons provide ways to graphically highlight critical content, steps to follow in acquiring content individually and with others, and ways to monitor progress and retention (Deshler & Schumaker, 2006). The four routines used in the study were unit organizers, question/exploration guides, cause and effect guides, and comparison (compare and contrast) tables. Examples of each of the routines, as well as terms and definitions used in this report, are presented in Appendix A.

The technology developed during the grant to use with the routines is a Google application called CORGI, which stands for Co-organize your learning. CORGI was designed as a Google application because this platform is free to use, and because the participating schools already use the platform with students. CORGI supports familiar Google functions including shared authoring and commenting, and it maintains the same graphic designs as the original paper-based routines. Several student supports are built into the application including embedded videos about how to use the routines, models of expert examples, text to speech and speech to text, and support for vocabulary and translation.

The combination of the routines and CORGI technology is therefore called *EU*. This *EU* project is the first to combine multiple research-based routines and technology as a means of teaching higher-order reasoning to secondary students.

DESIGN-BASED IMPLEMENTATION RESEARCH PROCESS AND DECISIONS

Leading up to this field test, SRI spent two years of intervention design following principles of Design-Based Implementation Research (DBIR) process (Penuel & Martin, 2015), including: engage, listen, revise, repeat. DBIR involves an iterative process where researchers and developers work with educators and students to design a product or intervention, pilot it, gather feedback, and use the feedback to drive product/intervention change.

The DBIR process for *EU* involved the completion of the following activities to engage teachers, administrators, students and researchers in the design process.

- Fifteen full-day or half-day researcher-practitioner design meetings

- Five student focus groups with middle school and high school students across both districts
- Seventeen student technology pilot focus groups
- Fourteen teacher interviews
- Three interviews with district administrators
- 54 class periods in which U.S. History and Biology teachers piloted the integrated units at multiple stages of development and shared their feedback via Google survey

During this process, it was revealed that the districts involved differed on multiple factors that influenced early design decisions. These differences included standards and curriculum alignment, technology policy, and district/school leadership culture. Because of these challenges, SRI adjusted the original schedule to choose to spend two years (instead of one) designing the innovation, and proposed a pilot study and a field study instead of a two-year field study.

DESCRIPTION OF PILOT STUDY AND SUBSEQUENT CHANGES

The pilot study, which was conducted by SRI, was designed as an RCT, and implemented in one district in California and one district in Virginia during the spring semester of the 2016-2017 school year. These districts had contributed to the design of the routines and technology. Teachers were blocked by school and subject, and randomly assigned to either *EU* or business-as-usual conditions in January 2017. Intervention teacher training started within two weeks after random assignment. Treatment teachers implemented *EU* in their U.S. History or biology classes using CORGI in spring 2017. Professional development modules were developed to train teachers over two days on how to implement *EU* with CORGI in whole class and small groups. In addition, teachers received coaching by request.

After completing the pilot study, the following changes were made for the field study.

- The field test was modified to include one practice unit and two study units. Coaching would continue during the practice unit, but will be a reduced number of hours during the study units
- One unit per subject (cells and Roaring 20s) were dropped from the field test
- The SCORE routine was dropped from the field test
- Preference to work in schools with at least four teacher participants (any combo of U.S. History/biology) due to issues of cost
- Teachers new to SIM were included
- Teachers on improvement plans were not included

More information on the DBIR process and changes from the pilot is included in Appendix B.

OVERVIEW OF THE FIELD TEST

The field test RCT conducted by Empirical Education, which is the focus of this report, was designed to address the following primary research questions.

- Did students in grades 9 through 12 who attended high school **science** classes that incorporated **science** *EU* demonstrate higher order content knowledge in the unit test scores in **science** compared to the scores of similar grades 9 through 12 students in high school biology classes that implemented business as usual (BAU) in spring 2018?

- Did 11th grade students who attended high school **social studies** classes that incorporated **social studies** *EU* demonstrate higher order content knowledge in the unit test scores in **social studies** compared to the scores of similar 11th grade students in high school U.S. History classes that implemented BAU in spring 2018?
- Did students in grades 9 through 12 who attended high school **science** classes that incorporated **science** *EU* and 11th grade students who attended high school **social studies** classes that incorporated **social studies** *EU*, as a group, demonstrate higher order content knowledge in their respective unit test scores compared to the scores of similar grades 9 through 12 students in high school classes that implemented BAU in spring 2018?

The secondary research questions include the following.

- Did special education students in grades 9 through 12 who attended high school **science** classes that incorporated **science** *EU* demonstrate higher **science** unit test scores compared to the scores of similar grades 9 through 12 special education students in high school **science** classes that implemented BAU in spring 2018?
- Did special education students in 11th grade who attended high school **social studies** classes that incorporated **social studies** *EU* demonstrate higher **social studies** unit test scores compared to the scores of similar 11th grade special education students in high school **social studies** classes that implemented BAU in spring 2018?
- Did special education students in grades 9 through 12 who attended high school **science** classes that incorporated **science** *EU* and 11th grade special education students who attended high school **social studies** classes that incorporated **social studies** *EU*, as a group, demonstrate higher order content in their respective unit test scores compared to the scores of similar grades 9 through 12 special education students in high school classes that implemented BAU in spring 2018?

In addition to these questions, which were identified prior to the study, we addressed several additional exploratory questions to better understand the results we obtained, including the following.

- Is there a difference in impact on student achievement depending on teachers' self-reported levels of comfort with technology?
- Is there a difference in impact on student achievement depending on biology content area, specifically, evolution compared to ecology?
- Is there a positive impact of *EU* on achievement by biology content area, or by history content area?
- What is the level of the treatment-control contrast in the use of SIM instructional practices deemed central to implementation of *EU*?
- Is there a correlation across randomized blocks between levels of fidelity of implementation and program impact?
- Is there evidence that *EU* had impact on instructional practices posited to mediate impacts on student achievement?

In addition to addressing these questions, this study documents the extent to which the core components of *EU* were implemented with fidelity. We will also provide descriptive results on classroom practices (as measured by teacher surveys) and contextual factors that support or hinder implementation (as described during teacher interviews).

For this experimental study, SRI recruited 13 teachers across five study schools in three districts to participate. Of the 13 teachers, seven taught biology and six taught U.S. History. All teachers were trained in implementing *EU* and received ongoing coaching during the school year. For each teacher, we randomly assigned each of their classes into one of two groups: the group using *EU (*or program group) or to the group who would continue with their existing instruction (control group), that is, "business as usual." We first paired each class with the one most similar and a random number

generator was used to determine which class in each pair would join the *EU* group and which class would be in the control group. Class rosters were established prior to random assignment. Implementation of *EU* occurred during the second semester of the school year and included three units in biology (genetics, evolution, and ecology), and three units in U.S. History (Great Depression, World War II, and Cold War).[1]

Methods

This section outlines the experimental design. Our experiment results in a comparison of outcomes for classes randomly assigned to *EU* and classes being taught using the district's current methods. The outcomes of interest are the student test scores on unit tests in biology and U.S. History. This section details the methods we used to assess the impact of *EU.* We begin with a description and rationale for the experimental design and go on to describe the program, the research sites, the sources of data, and the composition of the experimental groups.

EXPERIMENTAL DESIGN

How the Sample was Identified

How the participants for the study are chosen largely determines how widely the results can be generalized. The *EU* sample was one of convenience, chosen from school districts that were identified by using data from the Strategic Instructional Network. Districts or schools that were recruited had previously purchased SIM materials; had teachers trained in the use of some of the SIM routines and strategies; had a certified SIM trainer available; and identified a staff 'champion' for the proposed project. One district in California and one in Virginia agreed to participate in the pilot study and field test, and wrote letters of support for the proposal. When a third district was needed for the field test, the SRI project team reached out to the SIM network of certified trainers in Virginia, the location of one of the existing sites, to identify possible sites who met the above criteria. Empirical Education submitted a research application to the school board of the third district, and the study was approved.

Interested districts assigned a point of contact responsible for obtaining contact information for interested teachers and consent from district-level personnel. Eligible teachers would teach either biology in grades 9-12, or U.S. History in grade 11, and Empirical Education conducted informed consent webinars for interested teachers. Interested teachers then completed a consent form and baseline survey prior to receiving *EU* training. Four of the thirteen teachers that participated in the field test had previously participated in the pilot study, and one of those four had previously received *EU* training as part of the pilot. Students of study teachers (either *EU* or control) in one of the Virginia districts may have been exposed to SIM instructional practices through classes other than their target class they were enrolled in during the study. However, since an equal number of randomized classes in both conditions would have had this prior experience, exposure will be balanced between conditions.

Randomization

We would like to determine whether *EU* caused a difference in outcomes. To do so, we have to isolate its effect from all the other factors influencing performance. Randomization ensures that, on average, characteristics other than the

[1] The first unit (genetics and The Great Depression) were considered practice units and we did not include data from these units in the final analysis.

program that affect the outcome are evenly distributed between program and control groups. By evening out the effects of these factors between conditions, our estimate of the program effect is not confounded with effects of these other factors—technically it is "unbiased." Any remaining departures from the true values of the effects are due to chance differences between conditions.

There are various ways to randomize program participants to experimental conditions. Our research works within the organization of schools, not disrupting the existing hierarchy, in which students are grouped within sections that are nested under teachers in the schools. The level in the hierarchy at which we conduct the randomization is generally determined on the basis of the kind of program being tested. We attempt to identify the lowest level at which the program can be implemented without unduly disrupting normal processes or inviting sharing or "contamination" between control and program units. For example, school-wide reforms call for a school-level randomization while a professional development program that can be implemented individually per teacher can use a teacher-level randomization.

In the case of *EU*, we determined that the most appropriate and efficient unit of random assignment was the classroom level, based on the sample size of participating teachers. During the initial stages of work, Empirical Education discussed with SRI and CAST the potential limitations of having too small a sample of teachers ($n = 13$). Had randomization occurred at the teacher level, the study would have been too underpowered to detect a small to moderately-sized impact—ones typically found with educational interventions. Hence, Empirical Education implemented the design solution of randomizing classes within teachers, thereby increasing the sample of units randomized and the sensitivity to detect program effects. Appendix C provides our considerations for statistical power.

A concern with using a within-teacher randomized design is the potential for contamination. Because *EU* involved the technology component of CORGI, this partially mitigated concerns of contamination, as control classes were not given access to CORGI. However, we used additional strategies to minimize the risk of contamination. First, we explained to teacher participants—during a preliminary informational webinar—the importance, from a research perspective, of not using *EU* materials and techniques with their control classes, and we sent teachers email reminders of this point at the beginning of each unit. The Daily Implementation Logs also included a Yes/No question for teachers to self-report whether intentional or unintentional contamination may have occurred in that day's control classes. Finally, the end-of-study interviews asked teachers about their experiences in the treatment versus control conditions, to further capture the strength of the treatment-control contrast actually achieved, and to assess the extent of contamination, if any.

Because classes, instead of students, were assigned to the *EU* or control materials, this kind of experiment is often called a "group randomized trial." To increase the precision of our program impact estimates and increase design efficiency, we randomized classes within small blocks of similar classes. Using information from districts about student demographic and assessment data, as well as information from a baseline survey concerning characteristics of teachers' classes, for the most part we created matched pairs of classrooms within teachers. In a few exceptions, we used blocks with three sections, and in a couple of cases, we formed matched pairs across teachers.

Data informing block selection included proportions of low socioeconomic status, proportions of English Language Learners (ELLs), special characteristics (e.g. Advanced Placement, Honors, co-teacher), subject taught in the study class, and grade level. We also sought anecdotes from teachers about the similarities and differences between classes, a strategy that has proven effective for identifying similar classes in our past studies. In response, teachers offered comparisons of the overall sizes, energy levels, and manageability of their classes, as well as made notes on special

circumstances, such as whether the class had a co-teacher. Notes on special circumstances were particularly prioritized in the blocking process. This approach resulted in the same number of classes in each condition. All classes confirmed their rosters immediately, prior to randomization so that we had the most up-to-date baseline samples of classes and students. We used a random number generator to randomly assign one class within each pair to *EU*, and the other to control. The classes in each pair were assigned a random number drawn from a uniform distribution. The class with the higher number was assigned to treatment; the one other, to control.

The final configuration of blocks and classes in the experiment, as well as the random assignment status of each class, are displayed in Table 1 and Table 2 below (separated by biology and U.S. History). Overall, the study involved three districts, five schools, 13 teachers, 14 randomized blocks, and 30 classes (15 in each condition, with 18 in biology and 12 in U.S. History).

TABLE 1. UNITS IN THE EXPERIMENT: BIOLOGY SAMPLE

District	School	Biology teacher	Block	Condition 0 = Control; 1 = *EU*	Class
1	1	1	1	0	1
				0	2
				1	3
	2	2	2	0	4
				1	5
			3	0	6
				1	7
2	3	3	4	0	8
				1	9
	4	4	5	0	10
				1	11
		5	6	0	12
				1	13
3	5	6	7	0	14
				1	15
		7		1	16
			8	0	17
				1	18

TABLE 2. UNITS IN THE EXPERIMENT: U.S. HISTORY SAMPLE

District	School	U.S. History teacher	Block	Condition 0 = Control; 1 = *EU*	Class
2	3	8	9	0	19
				1	20
			10	0	21
		9		1	22
	4	10	11	1	23
		11		0	24
			12	0	25
				1	26
3	5	12	13	0	27
				1	28
		13	14	0	29
				1	30

What Factors May Moderate the Impact of *EU*?

The selected design allows us to measure the differential effectiveness of *EU* for specific subgroups of students, teachers, and units of instruction. These are variables that were measured before the experiment started, and that we had reason to believe would affect the magnitude of the effect of *EU*. Technically, these are called potential moderators because they may moderate (increase or decrease) the impact of *EU*. We measure the effect of the interaction between each potential moderator and the variable indicating assignment (i.e., to *EU* or control); that is, we measure whether the effect of *EU* changes across levels of each moderator.

For this study, we compared the program's effectiveness based on teachers' comfort level with using technology in the classroom, students' disability status, and among biology classes whether the unit of instruction was in Ecology or Evolution (we describe the rationale for this moderator analysis later as it is motivated by the main impact findings). We examined differential impact by teachers' levels of comfort with technology because *EU* is heavily technology-reliant. We expected that less impact on students would be conferred among teachers expressing less ease with using classroom technologies. We examined differential impact by student disability status because, as stated previously, *EU* is designed to improve student content learning and higher order reasoning, especially for students with disabilities or other learning challenges.

What Factors May Mediate Between *EU* and the Outcome?

We also examined impacts on a series of classroom practices that are potential mediators of the impact of *EU* on student outcomes. An impact on an intermediate instructional outcome means it may facilitate (mediate) impact on student achievement. While a lack of impact on an intermediate instructional outcome means the intermediate variable cannot mediate impact on student achievement.

Because of the small sample sizes in this study, we did not conduct formal mediation analyses; however, we examined impacts on specific intermediate variables. This allows us to rule out, with some degree of confidence, intermediate factors that are not mediators of impact on achievement. The analyses, and conclusions, should be considered exploratory, because of lack of power due to small samples (i.e., these analyses are capable of detecting only very large impacts, and there is a high probability of incorrectly concluding there is no impact on a mediator when there is a small but real effect). Because of this, we focused predominantly on the magnitudes of the effects.

SITE DESCRIPTION

The five study schools are spread across the two states, with three in Virginia and two in California, and nearly equally across the four National Center for Education Statistics (NCES) locale designations. Table 3 shows the school-level averages for the five study schools from publicly available NCES data.

TABLE 3. AVERAGE DEMOGRAPHICS OF STUDY SCHOOLS

Demographics	
Total schools	5
School characteristics	
Local designation: Rural	20%
Local designation: Town	20%
Local designation: Suburban	40%
Local designation: City/urban	20%
Total full-time equivalent teachers	347.6
Student characteristics	
Student to teacher ratio	16.96
Student population	5992
English language learners	8.1%
Students with IEPs[a]	11.7%
Low socioeconomic status	32.9%

TABLE 3. AVERAGE DEMOGRAPHICS OF STUDY SCHOOLS

Demographics	
Student Ethnicity	
White	52.4%
Black	12.2%
Hispanic	12.4%
Asian	17.2%
Pacific Islander	0.5%
American Indian/Native Alaskan	0.2%
Multi racial/No response	5.0%

Source: National Center for Education Statistics, 2017-2018 school year

Note. Percentages may not add up to 100% due to rounding of decimals

Data regarding the number of English language learners and students with IEPs was reported only at the district-level, not school-level.

To calculate the percentage of students with low socioeconomic status, the number of free lunch eligible and reduced-price lunch eligible students across the five schools was summed and divided by the total number of students at the five schools.

[a] IEPs are Individualized Education Programs

EU LOGIC MODEL AND OVERVIEW OF IMPLEMENTATION STUDY

During the 2016-2017 and 2017-2018 school years, the developers of *EU* worked to develop a program logic model and identify the key components of the intervention (Table 4 below). There are two key components: teachers receive sufficient support to implement the *EU*, and teachers increase their implementation of, adherence to, and quality of *EU* instructional practices, representing the inputs and outputs of the logic model. These are intended to impact teacher classroom use of SIM instructional practices, thereby increasing student collaboration and critical thinking in biology and U.S. History, which would in turn increase achievement on biology and U.S. History assessments, especially among special education students.

As a requirement of the National Evaluation of i3 (NEi3), we calculated fidelity of implementation (FOI) scores for each component of the *EU* program. The implementation study applies mixed methods to assess the key components of the logic model, including: presence of inputs, such as the delivery of PD and coaching by SRI/CAST; the usefulness of inputs measured through teacher surveys; and recorded levels of activities in terms of outputs, such as teachers' use of the routines, in terms of both frequency and quality, as well as student understanding and use of collaboration. We have assessed implementation fidelity in terms of the following components: (1) teachers receive sufficient support, which encompasses teacher attendance of PD and coaching, as well as their perceptions of the usefulness of the PD; and (2) teachers use of *EU* units, which includes teachers' adherence to, quality of, and reported usefulness of *EU* routines, as well as student self-reported understanding and use of collaboration.

TABLE 4. LOGIC MODEL FOR I3 *EU* STUDY

INPUT	PROXIMAL OUTPUTS	LONGER TERM OUTPUTS	TEACHER OUTCOMES	STUDENT OUTCOMES
Classroom level Biology and U.S. History teachers receive curricular materials for i3 *EU* **Key Component 1:** Biology and U.S. History teachers receive sufficient support: In-Person PD: Biology and U.S. History teachers receive sufficient support to use i3 *EU* materials by attending 3 days of PD Ongoing coaching: Biology and U.S. History teachers receive sufficient support by receiving at least 8 hours of coaching from SIM professional developers	Teacher **Key Component 2:** Teacher use of *EU*: Biology and U.S. History teachers use *EU*. Biology and U.S. History teachers implement one practice *EU* and two study *EU*s as per study design. Teachers deliver quality instruction, adhere to dosage, and report on likely effectiveness of the intervention on student performance. Student Students understand the purpose and application of the *EU*s in their biology and U.S. History classes	Teacher Improved implementation of, adherence to, and quality of *EU* instructional practices; improved effectiveness of *EU*	Teacher Improved/ increased implementation of SIM strategies (particularly the content enhancement routines specified in *EU*)	Student Improved achievement on end-of-unit content assessment measures

SCHEDULE OF MAJOR MILESTONES

The project began in September 2017 with the initiation of the contract and will end when the final report is delivered to SRI, CAST, and the NEi3. The study took place during the 2017-18 school year, with implementation occurring during the second semester of the school year. Table 5 presents an updated timeline of the major milestones of the study.

TABLE 5. MAJOR MILESTONES OF THE STUDY

Date	Milestone
October - December 2017	Recruitment of participating teachers
January - February 2018	Baseline data collection and roster identification
January - February 2018	Randomization
January - February 2018	Initiation of data collection activities for daily implementation logs
February 2018	Initiation of data collection activities for instructional practice surveys
May 2018	Final collection of daily implementation logs and instructional practice surveys
June 2018	Final collection of unit tests and student survey
April - June 2018	Collection of teacher interviews
November 2018	Submission of final report

DATA SOURCES AND COLLECTION

Teachers, schools, and districts provided the data to Empirical Education for this study. In addition to roster, demographic, and achievement data, we collected implementation data over the entire period of the experiment, beginning with the teacher trainings in November 2017, and ending with the schools' academic calendars in June 2018. We collected data through teacher background and instructional practice surveys, daily implementation logs, and teacher interviews to provide evidence of the implementation.

The precise schedule of teachers' implementation and relevant data collection varied by district and subject. The district's schedule of instruction and end of year assessments determined the start and end dates of each unit. In addition, whether or not the teacher had a block schedule determined the number of days of instruction in each unit. Teachers without block scheduling implemented units for about 15 instructional days each. Teachers with block scheduling implemented units for about 10 instructional days each. However, in both their *EU* and control classes, SRI asked teachers to implement on the same instructional schedule, in terms of number of days per unit, as well as content covered, with the only instructional difference being the use of *EU* themselves.

Table 6 outlines the overall timeline of the major data collection phases.

TABLE 6. DATA COLLECTION SCHEDULE FOR THE *EU* STUDY

	2017-2018 school year								
Data collection elements	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**April**	**May**	**June**
Baseline Survey		x	x	x					
District/school data request (for class rosters, demographic and prior achievement data)				x	x				
Training observations		x	x						
Daily implementation logs				x	x	x	x	x	
Instructional practice surveys					x	x	x	x	
Student end-of-unit assessment and end-of study survey					x	x	x	x	x
End-of-study teacher interviews							x	x	x

Source. Empirical Education staff

Next, we provide a description of each of the data collection instruments.

Teacher Baseline Survey

Using Empirical Education's SurveyCenter®, researchers administered a baseline survey to all participating teachers immediately after they agreed to participate in the study, and prior to the beginning of training and implementation. The survey was used to capture teachers' experiences and instruction prior to the program implementation. Teachers

had one week to complete the survey. Empirical Education staff downloaded the data for the survey each day, monitoring that the data was being submitted as intended, and following up with any non-respondents via email and phone.

The baseline survey asked teachers questions about their professional background and experience, level of education, and teaching certifications. In addition, it asked teachers to describe similarities and differences between their participating classes, as well as any special circumstances that classes may have. The latter information on classes was part of the data used in the blocking and randomization process.

Following questions about their professional experience and current classes, teachers answered how often they used SIM instructional practices. Additionally, they indicated how often they used practices such as lectures or presentations, explicit instruction, re-teaching, classroom discussions, advance organizers, and post organizers in their instruction. Moreover, the survey asked teachers how often they used practices such as games, hands-on experiences (i.e. labs, simulations), centers or stations, and non-linguistic representations (e.g., pictures, videos, graphs) to engage students in lessons. It also asked how frequently they used instructional practices to promote critical reading and thinking skills or supporting students in organizing and remembering content. Furthermore, it asked teachers how often they used various types of technology hardware (e.g., smartboards, overhead projectors, laptops, or tablets) and software (e.g., visual aids, electronic whiteboards). Teachers could answer these questions using a scale of "Never, Seldom, Sometimes, Often, Always."

For the last two sets questions, the survey asked teachers about their technology practices and self-efficacy. The first question showed teachers a set of statements regarding integrating technology into their classroom practices, such as "I can teach lessons that appropriately combine content, technologies, and instructional practices" and "I can select technologies to use in my classroom that enhance what I teach, how I teach, and what students learn." Teachers responded to the extent to which they agreed with the statements, according to a five-point Likert scale of "strongly disagree" to "strongly agree." These statements are adapted from the established Technological Pedagogical and Content Knowledge (TPAK) instrument, particularly the subsets of Technological Pedagogical Knowledge items (α = .93) and Technological Pedagogical Content Knowledge items (α = .89) (Schmidt, Baran, Thompson, Mishra, Koehler, & Shin, 2009). The final set of questions asked teachers about their beliefs in their ability in responding to things that can create difficulties in their instruction, such as "How well can you respond to difficult questions from your students?" and "How much can you do to adjust your lessons to the proper level for individual students?" according to a nine-point Likert scale from "nothing" to "a great deal." Researchers selected these statements from the established Teachers' Sense of Efficacy Scale longform instrument, particularly the subscale items on Efficacy in Instructional Strategies (α = .91) (Gibson & Dembo, 1984; Tschannen-Moran & Hoy, 2001).

District/School Data Requests

We requested and collected class rosters and student demographics from each of the school districts at the beginning of the study in January 2018, and collected updated rosters January through March 2018. These data are required to identify the baseline sample of students, match classes for randomization, conduct balance checks, and potentially serve as covariates in the impact analysis. Specifically, we asked the districts to provide the following student data.

- Name
- Unique identifier

- Grade
- Gender
- Ethnicity
- English proficiency status
- Disability status (whether or not student has a disability or is in special education, but not the specific condition)
- Socioeconomic status
- Classroom teacher name and unique identifier
- Course name and section
- School name
- Pretest scores

All student and teacher data with individually identifying characteristics were stripped of such identifiers for analysis, and the data were stored using security procedures consistent with the provisions of the Family Educational Rights and Privacy Act (FERPA). This experiment falls within the protocol approved by Empirical Education's Institutional Review Board (IRB), Ethical and Independent Review Services. Under this protocol and following FERPA guidelines, student or parental permission was not necessary, unless required by the district. Parental permission was required by, and obtained within, one of the participating districts.

Teacher Professional Development Observations

The *EU* professional development (PD) was divided into three sessions per school district, with district trainings held separately from one another. Empirical Education researchers observed the *EU* training in one district and received post-PD survey responses from teachers, from a survey administered by SRI. We coded the PD attendance and post-PD survey responses for adherence to the model as part of our implementation study.

Daily Implementation Logs

Using SurveyCenter, we administered daily implementation logs to all *EU* teachers during each day of implementation to assess adherence to *EU* instructional practices. Teachers could fill out the same log multiple times, identifying the specific date and period for which they were completing the log. During implementation, Empirical Education staff downloaded responses to the logs each morning, to determine whether each teacher had completed the necessary logs for the previous day, based on their schedule and number of *EU* classes. Empirical Education sent daily reminder emails to any teacher that had not yet completed the log for the previous day's classes. If there was still no response the following day, Empirical Education placed a phone call to the teacher.

On the daily implementation logs, teachers could answer yes or no to having used each of the different *EU* routines in their instruction that day. If they answered "yes" to any routine, there were several follow-up questions about their usage of the routine, such as the number of minutes spent on its use, whether they used CORGI with the routine, and if they co-constructed with students to develop the routine. Logs also asked teachers if they intentionally or

unintentionally used any of the *EU* routines in their most recent control classes. If so, they were to report the dates, control class periods, and routines used.

From January 29 through May 31, 2018, researchers maintained documentation of teachers' daily implementation schedules. Prior to the beginning of each unit, researchers emailed teachers to confirm whether their previously-stated implementation schedules were still applicable, or whether there had been unanticipated scheduling changes. At this time, we also reminded teachers to complete the daily implementation logs during unit implementation, as well as to not use the *EU* in their control classes, so as to maintain a contrast between the two conditions.

Instructional Practice Surveys

On the last day of implementation for each unit, Empirical Education administered an instructional practice survey. The survey asked the same set of questions for teachers' *EU* and control classes. However, for each survey, teachers randomly saw either the questions for their *EU* classes first, or the questions for control classes first, in order to avoid survey fatigue that may have created bias in responses. Each page of the survey reminded teachers in bold lettering as to which condition they should respond. Teachers had one week to complete the survey. Empirical Education staff downloaded the data for the survey each day, monitoring that the data was being submitted as intended, and following up with any non-respondents via email and phone.

This survey asked teachers the same set of instructional practice questions as the baseline survey, but, as described above, teachers responded about the study unit they had just completed, separately for the *EU* and control classes.

Student End-of-Unit Assessment and End-of-Study Survey

The end-of-unit assessments were written following general curriculum standards in Virginia and California, which teachers in both conditions were expected to teach. Certain items were created through consultation with assessment staff and experts in the general content area (university faculty, teachers).

The original design for the unit assessments was to have four items for each *EU* subcategory: two items that tap into higher-order thinking skills and two that do not. The assessments designed for the pilot study included extra items in each category so that items that did not perform well in the pilot could be removed from the assessments used in the field test. SRI computed Cronbach's alpha coefficients to determine the reliability for each unit test used in the pilot, and they analyzed item difficulty and item discrimination to select final items for the field test. The selected items had a difficulty range between .20 and .80 inclusively, and a discrimination of or above .20.

Before conducting the impact analyses in the field test, we examined specific psychometric properties of the assessments. Appendix D includes tables showing percent correct, point-biserial correlations, and response rates for individual items on the four unit tests. In Appendix D, we also show figures of the distribution of item difficulty (percent correct) values for each unit test.

The assessment performed well by conventional standards. There was no indication that students' rates of response were dropping off towards the ends of the assessment. Tests of internal consistency show Coefficient Alpha values of .91, .87, .86 and .76 for the Unit 2 and 3 biology and Unit 2 and 3 U.S. History assessments, respectively. The biserial correlations for most of the items were also considered high by conventional standards.

The final unit test form included student survey questions for students in *EU* classes. The student survey included three questions, including: satisfaction with the *EU* in helping them understand content, general satisfaction when compared to when the teacher did not use *EU*, and satisfaction with the CORGI technology.

The SRI data processing team developed, printed, and scored the end-of-unit assessments and end-of-study student survey. There was a careful firewall between SRI data processors and the SRI *EU* design team to avoid compromising the NEi3 requirement of independence. The SRI *EU* design team did not have access to any of the study data processed by the SRI data processing team. Additionally, Empirical Education only provided the SRI data processing team with class rosters for printing purposes. Empirical Education did not provide the design team with similar access. As a final precaution, rosters shared with the SRI data processing team contained dummy Study Student IDs, generated by Empirical Education, to be printed on the booklets.

At least three weeks in advance of anticipated administration of the end-of-unit assessments, Empirical Education securely shared class rosters with the SRI data processing team via our SecureServer®. These rosters included Study Student IDs, and no real student information, so as to maintain a firewall between SRI and Empirical Education. The unit title, school name, teacher name, and period number were also printed on the booklets.

The SRI data processing team printed the assessment booklets within one week, and returned them to Empirical Education at least two weeks in advance of anticipated administration. Booklets were immediately packaged and mailed to teachers, such that they arrived at least one week in advance of anticipated administration. In addition to the booklets, these packages included instructions for administration, a key linking Study Student IDs to real student names (depending on the privacy requirements of the district), instructions for return, and pre-addressed stamped return envelopes. FedEx and USPS were the preferred mail carriers to track all mailings and to limit the chance of loss.

Teachers returned the assessments to Empirical Education, who then returned them to the SRI data processing center for scoring. Following receipt of the scores, Empirical Education sent a list of student scores to each teacher. In order to further ensure independence, Empirical Education hand-scored a small random sample of assessments to verify that scoring is consistent across conditions.

End-of-Study Teacher Interviews

Approximately two weeks prior to the end of the third unit, researchers emailed teachers to assess their availability for a 30-minute interview. Empirical Education conducted the interviews using an online video conferencing channel, which recorded the interviews—with permission from teachers. Researchers transcribed the recorded interviews for further review.

The interviews sought to gain a better understanding of the following themes:

- Teachers' perspectives about instruction and implementation in their *EU* and Business as Usual classes;
- The amount of time they estimated spent teaching each unit in their *EU* and Business as Usual classes;
- Any difficulties faced implementing *EU*; and
- Any aspects of the program that may or may not be successfully scaled to other schools.

Sample questions include the following.

- Could you please walk me through a typical class that you lead using the *EU* materials?

- What were the main differences that you experienced between your *EU* classes versus your Business as Usual classes, if any?
- Would you recommend the *EU* program to other teachers? Why or why not?

After the interviews were completed, the research team reviewed the interview transcripts and coded them using qualitative methods. Both deductive and inductive codes were assigned to participant answers. In addition to coding the responses, the research team also counted each code's frequency across participants' answers; for instance, the number of participants who named a certain routine as being one that they used often in their *EU* classes.

Data Collection Response Rates

Table 7 includes the response rates for each data source.

TABLE 7. DATA COLLECTION RESPONSE RATES

Unit	Instrument	Expected	Received	Response rate
N/A	Consent and baseline surveys [a]	16	16	100%
1[e]	Daily implementation logs [b]	199	198	99.5%
	Instructional practice surveys [c]	13	13	100%
	Class sets of student end-of-unit assessments [d]	30	30	100%
2	Daily implementation logs [b]	137	137	100%
	Instructional practice surveys [c]	12	12	100%
	Class sets of student end-of-unit assessments [d, e]	30	30	100%
3	Daily implementation logs [b]	144	143	99.3%
	Instructional practice surveys [c]	12	12	100%
	Class sets of student end-of-unit assessments [d, e]	30	30	100%
	Teacher interview	12	12	100%

[a] Includes 3 co-teachers.

[b] Numbers for expected daily implementation logs are calculated based on the number of days of *EU* instruction each of the teachers has completed for the relevant unit.

[c] Numbers for expected instructional practice surveys are based on the number of teachers who have completed the relevant unit.

[d] Numbers for expected class sets of the end-of-unit assessments are based on the number of teachers who have completed the relevant unit.

[e] Includes numbers for teacher(s) who withdrew from implementation for Units 2 and 3, but who administered student assessments for Units 2 and 3.

FORMATION OF THE EXPERIMENTAL GROUPS

This section describes the study sample that we used to assess the impact of *EU*.

We start with the baseline sample, which consists of the participating classes that were randomly assigned to the *EU* or control group and for which we have information. The sample for which outcomes are analyzed may be modified somewhat from baseline through attrition or for other reasons that data become unavailable.

Baseline Sample

The *baseline sample* consists of the classes randomized to *EU* or control and includes all students in classes at the time of random assignment. The baseline student sample was determined prior to randomization and there were no joiners included in the analysis.

Ideally, by randomizing assignment into the two conditions, we create groups that are the same in terms of important characteristics – both those that are unobserved, as well as those that are observed, including demographics and prior achievement. In addition, because we randomized classes in blocks, we can expect a somewhat better balance on characteristics used to form the blocks. However, by chance, the groups are never exactly balanced and may differ on important characteristics that may be related to the outcome. Therefore, in this section we consider the equivalence of *EU* and control samples in terms of their baseline characteristics.

We consider the baseline samples in biology classes, U.S. History classes, and biology and U.S. History classes combined.

In Appendix E, we compare the composition of the control and *EU* groups in biology, U.S. History, and Combined (biology and U.S. History classes) at the point we received class rosters just prior to randomization (baseline sample). For each of the characteristics of this sample, we conducted a statistical test to determine the probability of observing a difference with magnitude as large as or larger than the one measured when in fact there is no difference.[2] While the randomization assures us that any imbalance was a result of chance, and is not an indication of selection bias, it is useful to examine the actual groups as formed at baseline to see whether the amount of imbalance is at a level we would expect to see less than 5% of the time (the standard conventionally used to assess if an effect is statistically significant). We show—in Appendix E—results of equivalence tests for baseline and analytic samples. None of the characteristics are distributed differently between conditions; that is, none of the differences reach statistical significance.

Attrition in this Experiment

We consider both overall and differential attrition. In a cluster randomized trial, we consider attrition of both the cluster randomized (i.e., sections) and the participants in those clusters (i.e., students).

Overall attrition. If the rate of overall attrition is large, even if there is no difference between conditions in the rate of attrition, then a loss of cases may induce bias in the result, if those who leave the program group are different from those who leave the control group. If overall attrition is above a certain level we must adjust for this difference in the analysis. For example, we would want to adjust for the effect of socioeconomic status if classes or students that attrite

[2] We used a *t*-test that adjusted for clustering of students in sections. The criterion for significance was set at <.05.

from the control group on average have lower socioeconomic status than classes or students that attrite from the *EU* group.

Differential attrition. If the rate of differential attrition is substantial, this also can induce bias in the result. If the rate of differential attrition is above a certain level, we must adjust for the characteristics that are imbalanced between conditions as a result of the differential loss of cases.

We report overall and differential attrition at the level of randomization and below. This allows calculation of potential for bias according to What Works Clearinghouse (WWC) standards (What Works Clearinghouse, 2017).

Among the 18 biology classes randomly assigned (nine to *EU* and nine to control) and 12 U.S. History classes randomly assigned (six to *EU* and six to control), none were lost to attrition during the course of the study. In other words, we obtained outcomes for one or more students present at baseline in each participating classroom. This leads us to consider attrition of students only.

The baseline sample of students includes all students present in study classes at the time of random assignment. Student counts from baseline to the analytic sample (for biology, U.S. History, and both subjects combined) are displayed in Table 8.

TABLE 8. COUNT OF POSTTEST IN BIOLOGY AND U.S. HISTORY SUBJECTS

	Count of students at baseline	Students with Unit 2 posttest	Students with Unit 3 posttest	Students with both Unit 2&3 posttest	Students with Unit 2 posttest only (and not Unit 3)	Students with Unit 3 posttest only (and not Unit 2)	Students with either Unit 2 OR Unit 3 posttest
Biology							
***EU* (N)**	194	170	174	163	7	11	181
Control (N)	219	198	198	186	12	12	210
Total N	413	368	372	349	19	23	391
U.S. History							
***EU* (N)**	111	105	107	103	2	4	109
Control (N)	128	122	120	115	7	5	127
Total N	239	227	227	218	9	9	236
Biology and U.S. History combined							
***EU* (N)**	305	275	281	266	9	15	290
Control (N)	347	320	318	301	19	17	337
Total N	652	595	599	567	28	32	627

In the case of *EU*, attrition was very low (as noted above, no classes attrited, and only a small number of students within classes attrited). The potential for results to be biased from non-equivalent samples is therefore low.

Table 9 (below) exhibit attrition calculations for outcomes within each unit, for each subject, and for the combined sample across subjects. We calculate attrition in two ways based on whether we consider attrition for: (1) students who do not have outcomes for both unit tests, and (2) students who do not have outcomes for either unit test. Under any approach to calculating attrition, the study is in the WWC category of having a "tolerable threat of bias under both optimistic and cautious assumption."

TABLE 9. ATTRITION COUNT OF POSTTEST IN BIOLOGY AND U.S. HISTORY SUBJECTS

	Count of students at baseline	Students with Unit 2 posttest	Attrition	Students with Unit 3 posttest	Attrition	Students with both Unit 2&3 posttest	Attrition (both outcomes)	Students with Unit 2 posttest only (and not Unit 3)	Students with Unit 3 posttest only (and not Unit 2)	Students with either Unit 2 OR Unit 3 posttest	Attrition (either outcome)
Biology											
***EU* (N)**	194	170	12.4%	174	10.3%	163	16.0%	7	11	181	6.7%
Control (N)	219	198	9.6%	198	9.6%	186	15.1%	12	12	210	4.1%
Total N	413	368		372		349		19	23	391	
Overall attrition			10.9%		9.9%		15.5%				5.3%
Differential attrition			2.8%		0.7%		0.9%				2.6%
Potential for bias			low		low		low				low
U.S. History											
***EU* (N)**	111	105	5.4%	107	3.6%	103	7.2%	2	4	109	1.8%
Control (N)	128	122	4.7%	120	6.3%	115	10.2%	7	5	127	0.8%
Total N	239	227		227		218		9	9	236	
Overall attrition			5.0%		5.0%		8.8%				1.3%
Differential attrition			0.7%		2.6%		2.9%				1.0%
Potential for bias			low		low		low				low

TABLE 9. ATTRITION COUNT OF POSTTEST IN BIOLOGY AND U.S. HISTORY SUBJECTS

	Count of students at baseline	Students with Unit 2 posttest	Attrition	Students with Unit 3 posttest	Attrition	Students with both Unit 2&3 posttest	Attrition (both outcomes)	Students with Unit 2 posttest only (and not Unit 3)	Students with Unit 3 posttest only (and not Unit 2)	Students with either Unit 2 OR Unit 3 posttest	Attrition (either outcome)
Biology and U.S. History											
***EU* (N)**	305	275	9.8%	281	7.9%	266	12.8%	9	15	290	4.9%
Control (N)	347	320	7.8%	318	8.4%	301	13.3%	19	17	337	2.9%
Total N	652	595		599		567		28	32	627	
Overall attrition			8.7%		8.1%		13.0%				3.8%
Differential attrition			2.1%		0.5%		0.5%				2.0%
Potential for bias			low		low		low				low

Analytical Samples: Equivalence of Participants following Attrition

Given the low attrition of participants, it is not necessary to examine equivalence of *EU* and control analytic samples, as the potential for bias in the impact estimates is low. However, for completeness, as with the baseline sample, we conducted a series of statistical tests to assess baseline equivalence for the analysis sample. Differences between conditions in distributions of baseline characteristics were not statistically significant for either the baseline or analytic samples (Appendix E).

ANALYSIS AND REPORTING ON THE IMPACT OF *EU*

Approach to Analysis

Before presenting the results, we discuss briefly the approach to analysis.

Average Impacts on Achievement Outcomes

We used a three-level hierarchical model to estimate impacts. Individual outcomes on unit tests are on the left side of the impact equation. On the right-hand side of the equation, at level 1 (occasion), we included a dummy variable that is coded 0 or 1 to indicate if the posttest was for a Unit 2 or Unit 3 outcome. At level 2 (student) we modeled a series of covariates including dummy variables indicating gender, ethnicity, English learner status, disability status, and grade level. For each covariate with missing values, we created a corresponding dummy variable to indicate cases with missing values (coded 1 if missing and 0 otherwise). The missing value of each covariate was zero. This "dummy variable imputation" approach is standard for addressing missing values for covariates in randomized experiments. Puma, Olsen, Bell, & Price (2009) found that the dummy variable approach to handling missing baseline covariates produces unbiased estimates of intervention effects in cluster randomized controlled trials. The method performed as well as multiple imputation and complete case analysis. We did not model pretests.[3] At level 3 (class), we included a dummy variable to indicate if the class was randomized to *EU* or control. At that level of analysis, we also added dummy variables to indicate class membership in randomization blocks. Random effects were modeled at each of the three levels of analysis: occasions, students and sections. We conducted impact analyses separately for biology and U.S.

[3] Prior to conducting the impact analysis, we decided not to include pretest as a covariate, because of the large number of different pretests administered to students, compounded by differences in the timing of pretest administration (e.g., prior spring or fall). The pretest differences occurred because the sample was located in multiple high schools in two different states. If pretest was included as a covariate, the analysis model would require dummy variables to account for the various combinations of pretest type and timing, and there would be relatively few students in each group. Partitioning the sample into these pretest groups would likely introduce more error into impact estimates, rather than improve precision. Although including pretest as a covariate in the impact analysis often improves the precision of the impact estimate, we anticipated that that would not be the case with so many different tests and timing.

A post-hoc analysis confirmed our decision. The set of baseline covariates that were included in the analysis model along with the dummy variable for treatment status accounted for 98% of class-level variance in outcomes (ICC was reduced from .251 for the unconditional (no fixed effects) model, to .005 when modeling the treatment effect with block and covariate effects), suggesting that including pretest as an additional covariate would offer little benefit.

History. For the combined analysis of impact across the two subjects, we used the same approach, but included a dummy variable for subject (biology or U.S. History) at level 3.

A further description of the impact model, including the equations, and the approach to handling missing data is described in Appendix F. Appendix G includes a description of how to interpret the effect sizes, estimates for fixed effects, and *p* values, as presented in the results section.

Sensitivity Analysis of Average Impacts on Achievement Outcomes

In addition to the main impact analyses, we report results of a series of sensitivity analyses designed to test if the main (benchmark) impact results can be replicated with other approaches to analysis. We assessed impacts (1) with the sample limited to students with posttest scores from both unit tests, (2) like the benchmark, but with randomization blocks modeled as random instead of fixed effects, (3) like the benchmark, but with posttest scores z-transformed within unit, and (4) by using the average of posttests across units as the outcome for individuals.

Moderator Analyses

These were limited to the combined sample, with one exception (to see whether impact for the unit on Evolution was greater than for the unit on Ecology). Generally, we estimate the moderating effect by including an interaction term in the statistical equation. This term multiplies together the variable that indicates whether the student is in treatment or control and the variable that indicates membership in the subgroups across which we want to assess the presence of a differential effect. The coefficient for this term measures the difference between the subgroups in the impact of the program. We assessed if impact varies by disability status, teachers' self-reported levels of comfort with technology based on the TPAK survey administered at baseline, and by biology content area (Ecology or Evolution), with the rationale for the last of these described in the "Results" section.

Treatment-Control Contrast

The study design involved randomizing classes within teachers (with two blocks identified across teachers). While we took precautions to limit the possibility of contamination and bias, we also assessed whether the four routines that constituted treatment were used in control classes. The four routines consisted of: (1) Use of unit organizers, (2) Use of question/exploration guides, (3) Use of cause and effect guides, and (4) Use of compare and contrast tables. Assessment of the difference between conditions in the use of these guides and routines was effectively a test of the treatment-control contrast. We expected large differences between treatment and control in the use of these routines. We did not consider other SIM routines or strategies as part of treatment, and we expected low contrasts on these other instructional practices. Teachers responded to their levels of use of each SIM instructional practice in treatment and control classes on two survey occasions, with response options: Never, Seldom, Sometimes, Often and Always. We examined the differences between treatment and control classes in their teachers' level of use of these practices.

Connections between Fidelity of Implementation and Impact

We investigated the relationship between fidelity of implementation (FOI) and impact. To do this, we examined the correlation between (1) block averages of *EU*-control differences in average student performance, and (2) block-level FOI. The rationale is that each of the 14 blocks represents a "mini-experiment" for which we can both estimate impact, and calculate levels of FOI. We can then look at the relationship between impact and FOI across the 14 blocks to see if

impact increases with increasing FOI. The number of these mini-experiments is small; therefore, the analysis will have low power and is strictly exploratory. Specifically, we look for positive trends in the correlations.

A limitation of the approach described here is the possibility of confounding of random sampling variation at the level of randomization (the class), with systematic variation in the treatment effect within each block. We are interested in estimating the latter, but for a given mini-experiment (i.e., per block), the treatment effect cannot be separated from class-level differences in outcomes that occur due to random variation in the sample; that is, by effects unrelated to the impact of *EU*.[4]

An indirect approach is to examine how much variation is present between clusters within matched pairs after adjusting for effects of covariates that account for random sampling variation. If, following adjustment, some variation in outcomes remains at the class level, it will reflect both random sampling variation not yet accounted for, and systematic differences across blocks in the treatment effect.

For exploration, we examined the correlation between block specific regression-adjusted estimates of the difference in outcomes between *EU* and control, and each of several fidelity scores. Positive correlations would indicate a relationship between fidelity and impact; while lack of a relationship could indicate either no relationship, or an underpowered test of the relationship. We are interested in whether there are any patterns in the correlation that are in the positive direction.

Impact on Mediators (for the Combined Sample and by Biology and U.S. History)

Mediation analysis is critical to understanding potential mechanisms through which impacts occur. Formal mediation analyses require large samples (Schochet, 2009) which this study did not afford. However, it is still instructive to assess if there are impacts on the posited mediators, since lack of impact on intermediate outcomes means they cannot mediate impact on important distal outcomes such as achievement.

To conduct this analysis given the samples available, we considered practices that SRI identified as important for facilitating impacts on achievement. They identified 17 practices that could be used in treatment or control. Teachers

[4] More formally, we can express the control mean in a randomized block as a quantity $Y_{ik}(C) = e_{ik}$ (for class *i* in block *k*), and the treatment mean in the same block as $Y_{jk}(T) = T_k + e_{jk}$ (for class *j* in block *k*.) In these expressions, e_{ik} and e_{jk} are random terms for the performance of each class, and T_k is the impact specific to block *k*. We would like to estimate T_k and e_{jk} separately. However, the problem we face is that we cannot de-confound these two effects. We can assume that $Var(e_{ik}) = Var(e_{jk})$, and we can estimate this quantity in the control group. We can also estimate $Var\left(Y_{jk}(T)\right) = Var(T_k + e_{jk})$, which allows us to estimate the variance in the treatment effect as $Var(T_k) = Var(T_k + e_{jk}) - Var(e_{ik})$ (assuming no correlation between the random error and systematic treatment effect terms.) We can then conduct a statistical test of the null hypothesis that $Var(T_k) = 0$. This does not yield block-specific values of impact to correlate with block-specific values of FOI; however, it would indicate if impact heterogeneity exists that may be correlated with FOI. A limitation is that if the variance in errors is much larger than the variance in impact, we would need a large sample to show that $Var(T_k + e_{jk}) > Var(e_{ik})$, (i.e., to conclude that impact heterogeneity is not zero). The variance in errors in this study is large, as it absorbs substantial between-block differences; therefore, we do not formally test if we can reject the null of no impact heterogeneity (the test is highly underpowered, and we would easily not reject the null hypothesis of no impact variation). We are limited to examining trends in correlations between regression-adjusted block-specific impact estimates, and block-specific FOI values.

were surveyed on their use of those practices at the end of each unit in both their treatment and control classes. Teachers were asked about the frequency of use of the following practices (each asked on a scale, Never, Seldom, Sometimes, Often and Always):

1. Explicit Instruction
2. Re-teach to a few students
3. Identifying similarities/differences (non-SIM)
4. Explicit strategy for asking clarifying questions (non-SIM)
5. Explicit summarizing strategy (non-SIM)
6. Explicit paraphrasing strategy (non-SIM)
7. Explicit vocabulary strategy (non-SIM)
8. Use of Graphic organizer (non-SIM)
9. Note-taking technique
10. Mnemonic device for remembering information
11. Rehearsing information aloud
12. Teacher laptop or Chromebook
13. Student laptop or Chromebook
14. Student tablet
15. Student collaboration on group and partner assignments
16. Teaching higher-order course content
17. Support for learners with different abilities

We were interested, first, in whether teachers reported greater use of certain practices in the treatment group, and second, whether greater impact was observed in biology or U.S. History.

For this analysis, we first averaged each teacher's response across surveys separately by condition (*EU* or control). This yielded two values in each randomized block for average frequency of use of each practice, one for treatment classes and one for control classes. Next, within each randomized block, we subtracted the control value from the treatment value. We then averaged this difference across the blocks; that is, within-block T-C differences were equally weighted and averaged for an overall mean difference in each practice. (Care should be taken in interpreting these values because averages were calculated over non-equal interval scales.)

Results

FIDELITY OF IMPLEMENTATION OF *EU*

As a requirement of the NEi3, we have calculated FOI scores for each key component of the *EU* program. The implementation study applies mixed methods to assess the key components of the logic model, including: presence of inputs, such as the delivery of PD and coaching by SRI/CAST; the usefulness of inputs measured through teacher surveys; and recorded levels of activities in terms of outputs, such as teachers' use of the routines, in terms of both frequency and quality, as well as student understanding and use of collaboration. We have assessed implementation fidelity in terms of the following key components: (1) Biology and U.S. History teachers receive sufficient support, and (2) Biology and U.S. History teachers use *EU*. The fidelity matrix (Table 10 and Table 11) provides an overview and fidelity thresholds for each indicator of the key components. Key Component 1 is made up of three indicators: (1) teachers received professional development (PD), (2) teachers received coaching, and (3) teachers found PD to be useful. These indicators were measured through PD attendance sheets, coaching logs, and a post-PD survey administered by SRI. Key Component 2 is made up of five indicators: (1) Adherence to *EU* implementation, (2) the quality of delivery of the *EU*, (3) the usefulness of the *EU* routines, (4) the extent to which students reported that the *EU* helped them understand the content, and (5) the extent to which students reported that the *EU* helped them collaborate with their classmates. These indicators were measured through the teacher daily implementation logs, instructional practice surveys, and the student survey.

Overall fidelity was not met for either key component. The following section reports separately on each indicator for the sample overall. Appendix H includes the complete FOI results by subject area.

Key Component 1: Biology and U.S. History Teachers Receive Sufficient Support

Table 10 provides an explanation of each indicator and results of fidelity of implementation for Key Component 1.

TABLE 10. FIDELITY MATRIX FOR THE *EU* KEY COMPONENT 1: BIOLOGY AND U.S. HISTORY TEACHERS RECEIVE SUFFICIENT SUPPORT

Indicator	Operational definition	Source of info/ data collection	Explanation of scoring	Fidelity threshold	Met fidelity?
Indicator 1. Teachers receive *EU* PD	PD attendance	Teacher sign-in sheet for 3 days of training; attendance records obtained from coaches	Teacher-level Attended entire 3-day training 0 = did not attend full training, 1 = attended entire training	If a teacher attends the entire 3-day training, he/she will get a score of 1.	12/13 (92%)
Indicator 2. Teachers receive coaching	Frequency and duration that teachers received ongoing coaching	Coach weekly log on *EU* implementation	Teacher-level Total hours receiving coaching 0 = < 8 hours 1 ≥ 8 hours	If a teacher receives ≥ 8 hours of coaching on *EU*, he/she will get a score of 1.	10/13 (77%)
Total teacher-level score for indicator 1 and 2			**1 = teacher attends entire 3-day training AND receives ≥ 8 hours of coaching. 0 = teacher does not receive a score of 1 on both indicators.**	**Total score = 1**	**9/13 (69%)**
Indicator 3. Teachers found PD to be useful	Usefulness of PD	Ten-item post-PD survey	District-level (out of 3 districts) Mean score of ratings on the post-PD survey by teachers in the district: 1=strongly disagree, 2=disagree, 3=agree, 4=strongly agree	The mean score on this survey is 3 or above.	2/3 (67%)
Criteria for implementing Component 1 with fidelity				At least 85% of teachers have a total score of 1 AND at least 2 of the 3 districts have a mean score of ≥ 3 on the post-PD survey.	Fidelity was not met.

The first indicator of Key Component 1, "Teachers received professional development," is defined as teachers receiving "initial professional development, make-up professional development, or follow-up professional development, as well as the duration of participation." Teachers were expected to attend three separate days of training offered within their respective districts. We collected attendance data through daily teacher sign-in sheets and attendance information from the instructional coaches (via email). At the teacher level, teachers received a score of 0 if they did not attend the entire training, and a score of 1 if they attended the entire training. Twelve out of thirteen teachers (92.3%) were present for all three days of training and, therefore, received a score of 1 on this indicator.

The second indicator, "Teachers received coaching," is defined in terms of the "Frequency and duration of ongoing coaching." Data for this indicator came from weekly logs kept by the *EU* implementation coaches within each district. Coaching was offered to teachers throughout the second semester (January through June) and included pre and post observations, lesson planning meetings, and lesson modeling. Sessions lasted between 10 and 60 minutes and targeted the core routines. Coaches would also note any implementation challenges and provide next steps for follow-up coaching.

At the teacher level, scoring for this indicator was determined by the total hours the teacher received coaching. The criterion scale for adequate implementation is: fewer than eight hours was not adequate; eight hours was adequate; greater than eight hours was exceeds adequate. If a teacher received greater than or equal to eight hours of coaching throughout the second semester, they received a score of 1. Ten of the thirteen teachers (77%) received eight or more hours of coaching, receiving a score of 1 for this indicator. The average number of coaching hours received for teachers who did not meet fidelity was 4.6, and the average number of hours received for teachers who did meet fidelity was 9.

At the teacher level, if a teacher attended all three days of PD (indicator 1) and received at least eight hours of coaching (indicator 2), the teacher received a total score of 1 for this key component. Nine out of the thirteen teachers (69%) met teacher-level fidelity for indicators 1 and 2.

We collected data for the third indicator, "Teachers found professional development to be useful," from the post-PD teacher survey. The survey included 10 items related to how satisfied the teachers were with the PD sessions (e.g., the PD made them enthusiastic about *EU* implementation, the presenters were well organized and informative, and that they can confidently use the *EU* with their students). Teachers responded to the questions on a four-point Likert scale from 1 = strongly disagree to 4 = strongly agree. Because these surveys were anonymized, there was no teacher-level scoring for this indicator. At the district level, scoring was determined by the mean score of ratings on the survey by teachers in the district. The fidelity threshold is a mean score of 3 or above on the surveys for a given district, for which the district received a score of 1. Overall, two out of the three districts received a score of 1 for this indicator.

At the component level, if at least 85% of all teachers had a total score of 1 for the two teacher-level indicators AND at least two of the three districts had a mean score of ≥ 3 on the teacher survey, then this key component was implemented with fidelity. Because only 69% of teachers received a total component score of 1, fidelity was not met for Key Component 1. As shown in Appendix H, five out of the seven (71%) Biology teachers and four out of the six (67%) US History teachers met teacher-level fidelity for indicators 1 and 2, indicating that neither group met fidelity for Key Component 1 separately.

Key Component 2: Biology and U.S. History Teachers Use *EU*

Table 11 provides an explanation of each indicator and results of fidelity of implementation for Key Component 2.

TABLE 11. FIDELITY MATRIX FOR THE *EU* KEY COMPONENT 2: BIOLOGY AND U.S. HISTORY TEACHERS USE *EU* UNITS

Indicator	Operational definition	Source of info/ data collection	Explanation of scoring	Fidelity threshold	Met fidelity?
Indicator 1. Adherence	Reported frequency of *EU* components used	Teacher Implementation Logs, Instructional Practice Survey	Teacher-level Reported use of: (a) Unit Organizer, (b) CORGI at least once with the Unit Organizer, (c) co-construction at least once with the Unit Organizer and once with each of the routines, (d) routines at least once each, (e) teaching background knowledge, (f) using "Cue-Do-Review" (see results below for exact scoring description). For teachers with more than one class, their points were averaged within teacher, across the two classes. In sum, teachers can earn 18 maximum points per unit.	85% (≥ 15.3) of the max possible points If a teacher received an average of at least 15.3 possible points averaged across the 2 study units, the teacher will get a score of 1	9/13 (69%)
Indicator 2. Quality of delivery	Reported combination of *EU* components used	Teacher Implementation Logs	Teacher-level Reported using combination of: (a) Unit Organizer + Corgi + Co-construct and (b) Routine + Corgi + Co-construct (see results below for exact scoring description). For teachers with more than one class, their points were averaged within teacher, across the two classes. In sum, teachers can receive a total of 12 points per unit.	85% (≥ 10.2) of the max possible points ≥ 10.2. If a teacher received an average of at least 10.2 points averaged across the 2 study units, the teacher will get a score of 1	6/13 (46%)

TABLE 11. FIDELITY MATRIX FOR THE *EU* KEY COMPONENT 2: BIOLOGY AND U.S. HISTORY TEACHERS USE *EU* UNITS

Indicator	Operational definition	Source of info/ data collection	Explanation of scoring	Fidelity threshold	Met fidelity?
Indicator 3. Usefulness	Usefulness of the *EU*	Instructional Practice Survey	Teacher-level 1 = not at all useful, 2 = less than moderately useful, 3 = moderately useful, 4 = more than moderately useful, 5 = very useful	If a teacher reported an average of 3 or above, he/she will get a score 1.	9/13 (69%)
Indicator 4. Student understanding	Students are satisfied that the *EU* helps them to understand the content.	Three-item question on Student Survey	Teacher-level 1=not satisfied, 2=just OK, 3=satisfied, 4=very satisfied; student scores are aggregated to the teacher level	If the average score of students' responses is ≥ 3, the teacher will get a score of 1.	8/13 (62%)
Indicator 5. Student collaboration	Students are satisfied that CORGI helps them to collaborate with their classmates.	Two-item question on Student Survey	Teacher-level 1=not satisfied, 2=just OK, 3=satisfied, 4=very satisfied; student scores are aggregated to the teacher level	If the average score of students' responses is ≥ 3, the teacher will get a score 1.	8/13 (62%)
Criteria for implementing Component 2 with fidelity				At least 85% of teachers have a total score of ≥ 4	5/13 (38%) Fidelity was not met.

The first indicator, "Adherence," is defined as the "Reported frequency of *EU* components used." We collected data for this indicator through the teacher daily implementation logs and the instructional practice survey. At the teacher-level, we determined scoring as follows, where teachers reported the following.

a. Using Unit Organizer in 100% of eligible classes (where eligible excludes the first day, last day, and 1 extra day). Teachers would receive a maximum of 3 points per unit, with partial credit granted.

b. Using CORGI technology at least once with the Unit Organizer and once each with the three routines (question exploration, cause and effect, and compare/contrast), in four separate class periods. One point was granted per paired use, for a maximum of 4 points per unit, with partial credit granted.

c. Using co-construction at least once with the unit organizer and once with each of the routines. One point was granted per paired use, for a maximum of 4 points per unit, with partial credit granted.

d. Using routines (question exploration, cause and effect, and compare/contrast) at least once each. One point was granted per use for a maximum of 3 points per unit, with partial credit granted.

e. Teaching background knowledge in at least 3 class periods. One point was granted per period for a maximum of 3 points per unit, with partial credit granted.

f. Using "Cue-Do-Review" "Most or Always" on the teacher instructional practice survey. One point was granted per unit, but no partial credit possible.

For teachers with more than one class, their points were averaged across classes. In sum, teachers could earn 18 maximum points per unit. If a teacher received an average of 85% (15.3) or above of the total possible points across the two study units, the teacher received a score of 1. Overall, nine out of thirteen teachers (69%) received a score of 1 on this indicator. The average score across the two units was 14.7, with the same average score for Unit 2 and Unit 3. The average overall score (i.e., across the two units) for the nine teachers who met fidelity was 16.8. The average score for the four teachers who did not meet fidelity (including one teacher who did not implement either study unit) was 9.9.

The second indicator, "Quality," is defined as the "Reported combination of *EU* components used" and teacher daily implementation logs provided the data. At the teacher-level, the following reported by teachers determined scoring.

a. Using the Unit Organizer + Corgi + Co-construction, during one class period on the same day. Teachers received 1 point per element for a maximum of 3 points per unit.

b. Using a Routine + Corgi + Co-construction together. Teachers received 1 point per element, during at least 3 class periods, for a maximum of 9 points per unit.

In sum, teachers could receive a total of 12 points per unit. If a teacher received an average of 85% (10.2) or above of the possible points between the two study units, the teacher received a score of 1. Overall, six out of thirteen teachers (46%) received a score of 1 on this indicator. The average overall score (i.e., across the two units) for the six teachers who met fidelity was 11.75 (with five of the six earning the maximum of 12 points on both units). The average score for the seven teachers who did not meet fidelity (including one teacher who did not implement either study unit) was 6.2.

The teacher instructional practice survey provided data for the third indicator, "usefulness of the *EU*." On each of the unit surveys, teachers reported on the usefulness of the four core *EU* routines (Unit Organizer, Question Exploration,

Compare and Contrast, and Cause and Effect) on a five-point Likert scale, from 1 = not at all useful to 5 = very useful. If a teacher reported an average of 3 or above on this scale across each routine, he/she received a score of 1. Overall, nine out of thirteen teachers (69%) received a score of 1 on this indicator. (In the *Descriptive Findings* section below, we provide more information about the reported level of usefulness of each *EU* core routine.)

Data for the fourth indicator, "student understanding," and fifth indicator, "student use of collaboration," are from the student survey administered at the same time as the Unit 3 assessment. For "student understanding," students reported how satisfied they were that the unit organizer, routines, graphics, and CORGI technology helped them:

a. understand how to think about the important topics in the units,

b. focus your attention on what was important to learn in the units, and

c. study for tests.

For "student use of collaboration," students reported how satisfied they were that CORGI technology helped:

d. you to work with your classmates on completing unit organizer or routine, and

e. everyone in the class participates while working on completing a unit organizer or routine.

Students responded to each question on a four-point Likert scale of 1 = not satisfied, 2 = just OK, 3 = satisfied and 4 = very satisfied. Overall, eight out of thirteen teachers (61.5%) received a score of 1 on each of these indicators.

At the component level, if the fidelity total score for a teacher was four or above, the teacher implemented *EU* with fidelity. Overall, five out of thirteen teachers (38.4%) received four or more points. Therefore, fidelity was not met for Key Component 2. As shown in Appendix H, neither Biology nor US History teachers met fidelity separately, as three out of seven (43%) Biology teachers and two out of six (33%) US History teachers received four or more points.

DESCRIPTIVE FINDINGS RELATED TO *EU* IMPLEMENTATION

In this section, we provide descriptive statistics on teacher-reported implementation and usefulness of the SIM instructional practices in their *EU* classes.

Frequency of Use of SIM Instructional Practices

On each instructional practice survey, teachers were asked how often they used the SIM instructional practices in their *EU* classes when teaching the most recently completed unit. Responding on a five-point Likert scale of "always" to "never" teachers reported about the following practices.

1. SIM Course Organizer Routine
2. SIM Unit Organizer Routine
3. SIM Compare Contrast Routine
4. SIM Question Exploration Routine
5. SIM Cause and Effect Routine
6. SIM LINCS Vocabulary Strategy
7. SIM Paraphrasing Strategy
8. SIM Self-Questioning Strategy

9. SIM Summarization Strategy
10. SIM Concept Mastery
11. Use of SIM "Cue-Do-Review"

Figures 1 and 2 show the distribution of teachers' responses to the question on the Units 2 and 3 surveys.[5,6]

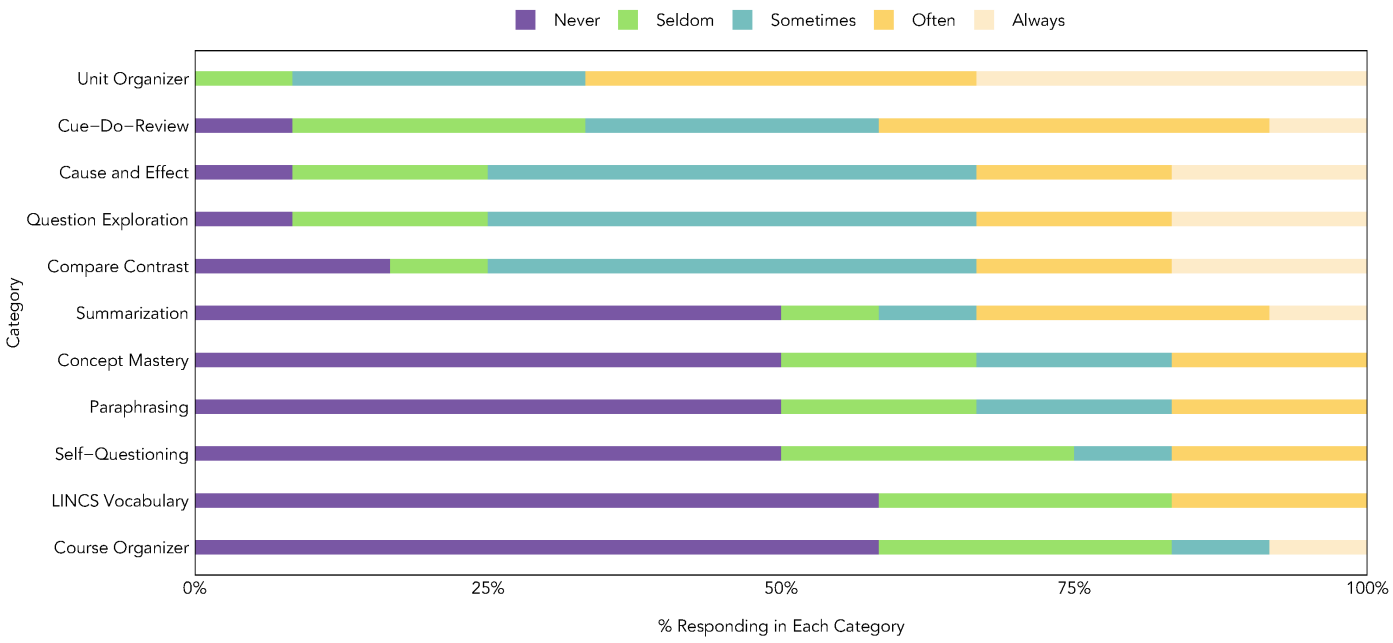


FIGURE 1. FREQUENCY OF USAGE OF SIM INSTRUCTIONAL PRACTICES IN TREATMENT CLASSES, UNIT 2 INSTRUCTIONAL PRACTICE SURVEY

[5] Because Unit 1 was a "practice unit" in implementing *EU* for the teachers, we are only analyzing the survey responses for Units 2 and 3 in this report.

[6] The bars have been ordered by the most to least frequently used practices based on the percentage of teachers responding "always" or "often".

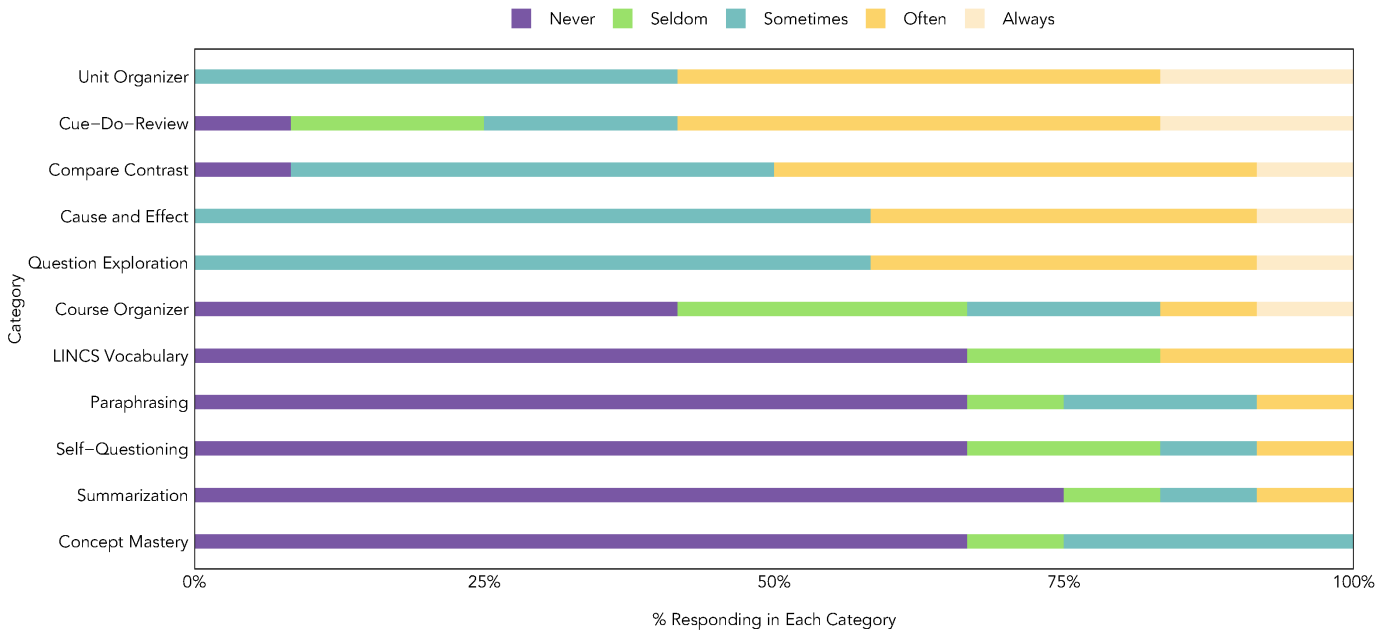


FIGURE 2. FREQUENCY OF USAGE OF SIM INSTRUCTIONAL PRACTICES IN TREATMENT CLASSES, UNIT 3 INSTRUCTIONAL PRACTICE SURVEY

Based on responses to both surveys, the five most frequently used routines were the Unit Organizer, Cue-Do-Review, Question Exploration, Cause and Effect, and Compare and Contrast routines. This result is expected given that these routines were the ones that the teachers were explicitly instructed to use during the PD (except for Cue-Do-Review). Furthermore, the teacher interviews corroborated this pattern. During the interviews, when teachers answered what learning routines they used most often in their *EU* classes, the four most frequently mentioned routines were the core routines: Compare and Contrast, Unit Organizer, Cause and Effect, and Question Exploration. Teachers often used the Unit Organizer to introduce the lesson or class content; some also used it to review material from the previous class. Many teachers reported that they liked using Compare and Contrast because they felt that it was the easiest routine for students to understand and also because of its interactive nature.

Additionally, Figures 1 and 2 show that between the Unit 2 and Unit 3 surveys, the proportion of teachers answering "Never" and "Seldom" to how frequently they used the Unit Organizer, Cue-Do-Review, Question Exploration, Cause and Effect, and Compare and Contrast routines decreased, while the proportion of teachers answering "Often" and "Sometimes" increased. This may suggest that the instructional coaching that teachers received in using those routines was effective or that the teacher became more comfortable with those routines, insofar that teachers began to implement the routines more regularly in their instruction.

Usefulness of *EU* Core Routines

On each instructional practice survey, we asked teachers how useful they found the core SIM instructional practices when teaching the most recently completed unit in their *EU* class. Responding on a five-point Likert scale of "very useful" to "not at all useful," we asked teachers to respond about the usefulness of the following routines.

1. SIM Unit Organizer Routine
2. SIM Compare and Contrast Routine
3. SIM Question Exploration Routine
4. SIM Cause and Effect Routine

Figures 3 and 4 show how teachers responded regarding the usefulness of each routine in their *EU* class for Unit 2 and Unit 3, respectively.[7]

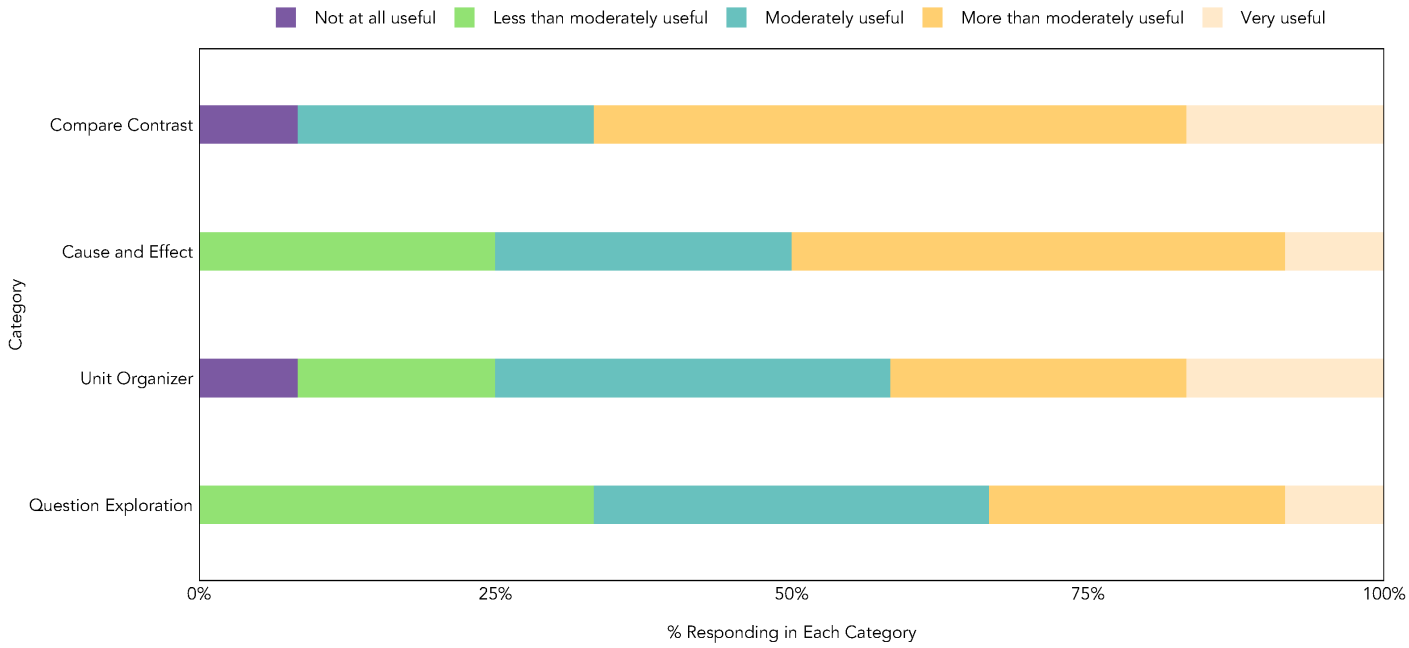


FIGURE 3. USEFULNESS OF SIM INSTRUCTIONAL PRACTICES, UNIT 2 INSTRUCTIONAL PRACTICE SURVEY

[7] The bars have been ordered by the most to least useful routine based on the percentage of teachers responding "very useful" or "more than moderately useful.

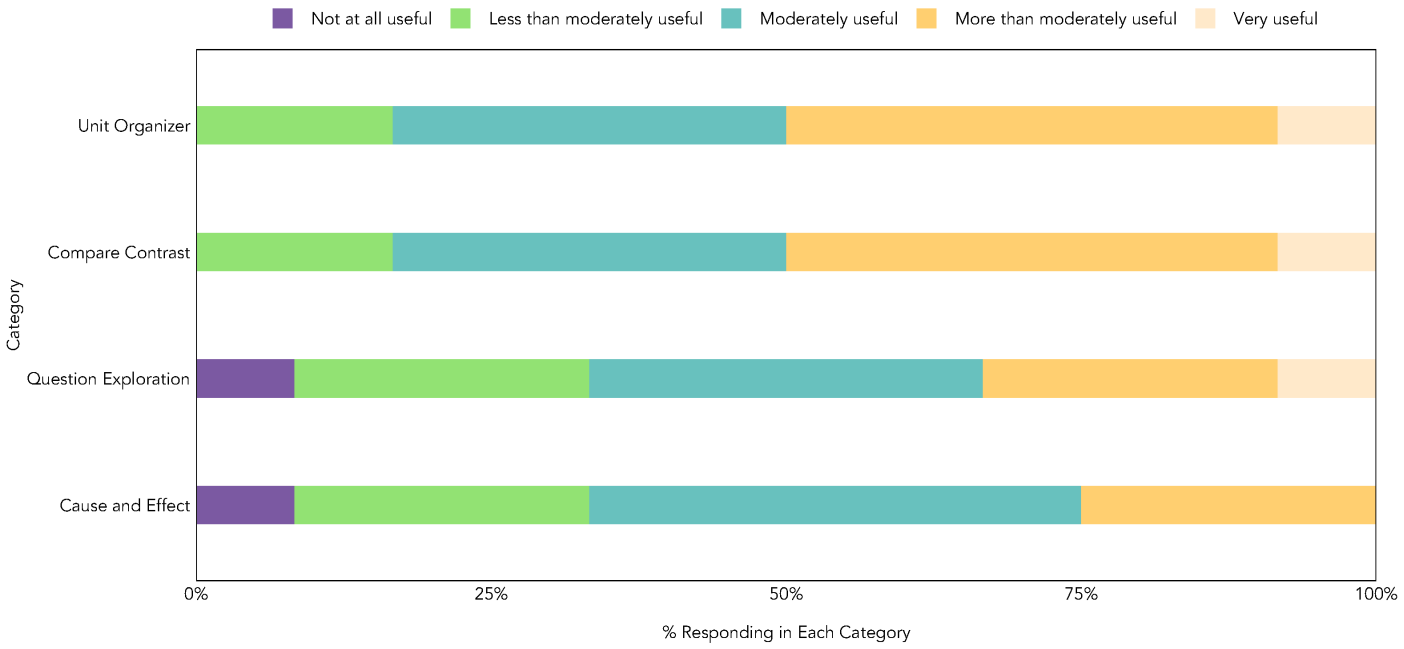


FIGURE 4. USEFULNESS OF SIM INSTRUCTIONAL PRACTICES, UNIT 3 INSTRUCTIONAL PRACTICE SURVEY

The two graphs show that teachers' perceptions of the routines' usefulness varied between Units 2 and 3. To better understand how teachers felt about the routines' utility, we asked teachers the following questions during the interviews.

- What learning routines do you find most helpful for struggling learners in particular? Why do you think these are helpful for them?
- Which *EU* learning routines did you find to be the most useful? (E.g., Unit Organizer, Compare and Contrast, Cause and Effect, and Question Exploration) Please describe.
- Which *EU* learning routines did you find to be the least useful? (E.g., Unit Organizer, Compare and Contrast, Cause and Effect, and Question Exploration) Please describe.

While there were some patterns in how teachers responded to the questions about which routines were useful or not, teachers' perspectives on each of the routines were not monolithic. Some teachers viewed some routines favorably, while others were less favorable towards the same routine.

For struggling learners, Unit Organizer, Compare and Contrast, Cause and Effect, and Question Exploration were all cited at least once across all interviewees as being helpful. However, the Cause and Effect and Compare and Contrast routines were most frequently cited. Teachers cited Cause and Effect because of its straightforward delivery. For instance, one teacher stated that the Cause and Effect routine was "the most linear. This is what we're doing, this caused it, this is what happened, use that to answer the question."

Additionally, eight out of twelve teachers interviewed cited the Compare and Contrast routine for a variety of reasons, such as how it encouraged students to break questions down into smaller questions, its focus on building vocabulary,

and because students were familiar with and enjoyed the practice of identifying similarities and differences. One biology teacher noted that Compare and Contrast seemed:

> "…helpful for [English Learners] to see what is it exactly we are comparing because it takes it a step further…what are these categories specifically that we're comparing? It kind of lumps them together in a way that they think of more generalized terms rather than more specific terms so they can get caught up on the specifics of whether something is made of carbon or whether something is not made of carbon, for example…I feel like the compare/contrast, they're probably used to seeing a Venn diagram, but in this way also able to see next step, higher order thinking about 'Hey, those are categories'."

Similarly, all interviewees cited all four *EU* routines at least once as being most useful overall—with Compare and Contrast cited most frequently (8 out of 12 teachers interviewed). Several teachers mentioned that they felt this routine was easiest for students to understand, and they also liked that the routine could be used with graphic aids such as Venn diagrams and compare-contrast charts.

In response to the question asking which routines they felt were *least* useful overall, teachers cited all routines except for Compare and Contrast. They most frequently cited Question Exploration as least useful (6 out of 12 teachers interviewed). Several teachers felt that this routine lacked sufficient scaffolding to function as a device for higher-order thinking, and described having to closely guide students through it. Consequently, they felt that this inhibited students from co-constructing. For instance, as one teacher described,

> "I couldn't cut them loose and say, 'Let's work on this and sync up after.' They just sat there and looked at me, and this was my advanced class. So when the routine was finally filled out, it was so wordy and so full of information, that I think the visual aspect was off. I think the effectiveness decreased as there was way too much information."

STUDENT-LEVEL IMPACT RESULTS

We analyzed impacts of *EU* on (1) biology, (2) U.S. History, and (3) biology and U.S. History combined. As described previously, we assessed outcomes using developer-produced measures with adequate internal consistency reliability. There were two unit tests for biology and two unit tests for U.S. History used in the impact analysis. Students were scored in terms of the percent correct metric on up to two unit tests. We analyzed outcomes as repeated measures (i.e., performance on up to two unit tests) nested within individuals.

Benchmark Impact Results

We estimated regression-adjusted average impacts of *EU* on biology, U.S. History, and general achievement (biology and U.S. History considered together). We applied hierarchical linear models and repeated measures analysis using both SAS and HLM software.

Our approach was to analyze impact by sequentially adding effects until we arrived at the a-priori specified full-covariate model. We summarize benchmark impact results in Table 12 below. (Full estimates corresponding to the impact models are reported in Appendix I.)

We observe no impact for biology, with a covariate-adjusted standardized effect size of .01, (p = .892). We observe a positive impact for U.S. History, with a covariate-adjusted standardized effect size of .32, (p = .037). Combining the biology and U.S. History samples, we estimated an effect size of .14 (p = .067).

TABLE 12. RESULTS FOR BIOLOGY AND U.S. HISTORY OUTCOME IMPACT ANALYSIS

	Condition	Means	Standard deviations[a]	No. of posttest scores	No. of students	No. of teachers	Effect size	*p* value	Change in percentile ranking
Biology									
Unadjusted effect size[a]	Control	70.77	22.45	396	210	9	0.01	.958	0%
	EU	71.00	22.19	344	181	9			
Adjusted effect size[b]	Control	70.77					0.01	.892	0%
	EU	71.02							
U.S. History									
Unadjusted effect size[a]	Control	49.39	22.16	242	127	6	0.33	.214	12%
	EU	56.51	20.40	212	109	6			
Adjusted effect size[b]	Control	49.39					0.32	.037	12%
	EU	56.18							
Biology and U.S. History combined									
Unadjusted effect size[a]	Control	62.66	21.97	638	337	13	0.14	.516	6%
	EU	65.65	21.53	556	290	13			
Adjusted effect size[b]	Control	62.66					0.14	.067	6%
	EU	65.77							

[a] The unadjusted effect size is the regression-adjusted impact estimate in a model without covariates divided by the pooled standard deviation in outcomes.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Note. Full estimates corresponding to the impact models are reported in Appendix I.

Source. Empirical Education staff calculations

Sensitivity Analyses

We conducted a series of sensitivity analyses. The results are displayed in Table 13 through Table 15 below. We assessed impacts (1) with the sample limited to students with posttest scores from both unit tests, (2) like the

benchmark, but with randomized blocks modeled as random instead of fixed effects, (3) like the benchmark, but with posttest scores z-transformed within unit, and (4) by using the average of posttest across units as the outcome for individuals.

We observe that for U.S. History and Combined outcomes the results are robust in terms of their magnitudes; however, for U.S. History, the *p* values fluctuate around significance level .05.

TABLE 13. SENSITIVITY ANALYSES FOR IMPACTS ON BIOLOGY

Model	Model Variant	Standardized Effect Size	Number of students in *EU* and control
Benchmark		ES=0.01	N(total)=391 N(T)=181 N(C)=210
SA 1	Like the benchmark but limited to students with both unit tests	ES=-0.04	N(tot)=349 N(T)=163 N(C)=186
SA 2	Like the benchmark but with blocks modeled as random	ES=-.07	N(total)=391 N(T)=181 N(C)=210
SA 3	Like the benchmark but with posttest scores z-transformed within unit	ES=0.01	N(total)=391 N(T)=181 N(C)=210
SA 4	Posttests are averaged across units for students with both unit test scores	ES=-0.04	N(total)=349 N(T)=163 N(C)=186

TABLE 14. SENSITIVITY ANALYSES FOR IMPACTS ON U.S. HISTORY

Model	Model Variant	Standardized Effect Size	Number of students in *EU* and control
Benchmark		ES=0.32	N(total)=236 N(T)=109 N(C)=127
SA 1	Like the benchmark but limited to students with both unit tests	ES=0.29	N(total)=218 N(T)=103 N(C)=115
SA 2	Like the benchmark but with blocks modeled as random	ES=0.32	N(total)=236 N(T)=109 N(C)=127
SA 3	Like the benchmark but with posttest scores z-transformed within unit	ES=0.32	N(tot)=236 N(T)=109 N(C)=127
SA 4	Posttests are averaged across units for students with both unit test scores	ES=0.32	N(total)=218 N(T)=103 N(C)=115

TABLE 15. SENSITIVITY ANALYSES FOR IMPACTS ON BIOLOGY AND U.S. HISTORY COMBINED

Model	Model Variant	Standardized Effect Size	Number of students in *EU* and control
Benchmark		ES=0.14	N(total)=627 N(T)=290 N(C)=337
SA 1	Like the benchmark but limited to students with both unit tests	ES=0.11	N(total)=567 N(T)=266 N(C)=301
SA 2	Like the benchmark but with blocks modeled as random	ES=0.12	N(total)=627 N(T)=290 N(C)=337
SA 3	Like the benchmark but with posttest scores z-transformed within unit	ES=0.14	N(total)=627 N(T)=290 N(C)=337
SA 4	Posttests are averaged across units for students with both unit test scores	ES=0.11	N(total)=567 N(T)=266 N(C)=301

TABLE 15. SENSITIVITY ANALYSES FOR IMPACTS ON BIOLOGY AND U.S. HISTORY COMBINED

Model	Model Variant	Standardized Effect Size	Number of students in *EU* and control

Moderator Analyses (Analyzed on the Combined Sample)

Disability

We assessed whether impact varies depending on a student's disability status. In total, 499 students had information about disability status and could be included in the analysis. Among the 111 students designated as disabled by the district datasets, 41 students had an unspecified disability, 21 were designated as having a specific learning disability, 13 had other health impairments, 3 had a speech or language impairment, 1 student was designated autistic, and 1 student was designated as having an emotional disturbance. Thirty-one students were designated "504" (these are students who are recommended for assessment for a disability and qualify for certain accommodations, but who do not qualify for an Individualized Education Program.)

We found a positive differential impact of *EU* on achievement depending on disability status. The added value to impact is 8.37 units in the percent correct metric (t=2.15, p=.040). Impacts for each subgroup were 1.472 for students without disabilities (t=0.87, p=.405) and 9.544 for students with disabilities (t=2.78, p=.017), both reported in terms of the percent correct metric. Disabled students in the control condition scored 47.43 scale score units at posttest, while non-disabled students in the control condition scored 66.91 at posttest. With the differential impact reported above, we expect disabled students to achieve an average posttest score of 56.97, and non-disabled students to score 68.38, in the *EU* condition.

Facility with Technology from Baseline Survey

We assessed whether impact varies depending on a teachers' incoming score on the 7-item TPAK-adapted survey. Each item asks teachers to indicate level of agreement on a five-point Likert scale with response options ranging from strongly disagree to strongly agree. Based on responses of the 13 teachers in the study, the standardized Cronbach Alpha was .90. We found no differential impact of *EU* on achievement depending on level of baseline score on the TPAK-adapted survey. The added value to impact with a one-unit increase in TPAK-adapted score was 2.06 units in the percent correct metric (t=.713, p=.488). This result is based on a weakly-powered analysis given that we had only 13 teachers (and therefore 13 scores on the TPAK-adapted) across which to differentiate impact.

Additional Exploratory Analysis of Difference in Impact by Content Area

Primary Results and Related Hypotheses

Based on the initial findings, the program developers offered the following hypotheses of why a positive impact was observed in U.S. History but not biology, as well as theory to predict a difference in the impact *within* biology units. We report the conjectures and additional tests in this section.

To explain the observed difference in impact between U.S. History and biology, the developers noted this hypothesis:

> Each enhanced unit in U.S. History focused on a single topic (World War II and the Cold War). Within both U.S. History Enhanced Units, the CERs (routines) expanded understanding and reasoning about concepts, events, and actions associated with the topics of World War II and the Cold War. As a result, there was instructional time to develop adequate background knowledge on each topic and to engage students in higher reasoning and deep understanding.
>
> The developers proposed that the content of enhanced units best support student learning when they focus on a single topic, allow adequate time, and use instructional supports that all relate to the critical topic of the unit and build sequential understanding.
>
> To verify this hypothesis the developers stressed the importance of testing the finding of a "need for closely aligned and scaffolded …sequences of the routines." That is, it was hypothesized by developers that EU will work best with logically sequenced content. The developers then made the a-priori hypothesis that, *within* biology, the unit on Evolution proceeds logically with one topic building from the last—from foundational theories of evolution, to extended exploration of the theory of natural selection to explain evolution, and then to how natural selection leads to selection of traits. In contrast, material addressing the other unit—Ecology—is less logically sequenced and does not allow for sequential and scaffolded understandings built incrementally from one CER to the next.

More specifically, the developers stated the following.

> **In the Evolution Unit**, the focus is on a critical topic: natural selection. Adequate background knowledge was developed and in-depth exploration achieved with the following: (a) comparing and contrasting the different theories of natural selection of Darwin and LaMarck with a Comparison Routine, (b) exploring the multiple types of evidence that support evolutionary change due to natural selection, such as fossil findings, with a Question Exploration Routine, and (c) developing an understanding of how organisms with the most favorable heritable traits will survive and reproduce with a Cause-and-Effect Routine.
>
> **By contrast, in the Ecology Unit**, more than one complex subtopic was explored for each of the three CERs (routines): (a) biotic and abiotic factors taught with a Comparison Routine, (b) the Carbon Cycle taught with the Question Exploration Routine, and (c) the Effect of Biomass at Each Tropic Level on available energy for use by organisms taught with a Cause-and-Effect Routine. As a result, even more complex conceptual understanding needs were included: ecosystem, biosphere, food pyramid (taught in some schools as pyramid of numbers), tropic level, producer, consumer, etc.

Results of Tests of Exploratory Hypotheses

After reporting the results of the impact of *EU* on U.S. History and biology, and given the developers' prediction concerning expected impacts within biology, we set out to substantiate their a-priori hypothesis.

To estimate the difference between Ecology and Evolution in the impact, we first z-transformed scores within each of the units. We then specified a model with outcomes for both tests modeled at level 1, with an indicator of which unit the test outcome is for (Evolution or Ecology) at that level. We modeled outcomes within students, nested in clusters (section), with dummy variables for matched pairs (much like the benchmark impact model). However, we also set the

Evolution-Ecology score gradient to vary randomly across units randomized (which allowed a more conservative test) and modeled a cross-level interaction between treatment status (*EU* or control) and biology unit (Evolution or Ecology).

Results of our analysis, to an extent, were consistent with the program developer expectation based on their rationale that is described above. Specifically, we found that students on average experienced greater impact of *EU* on assessment of Evolution than on Ecology. The difference was .171 standardized effect size units (t=2.00) and was marginally statistically significant (p=.063) (i.e., we could have moderate confidence in there being a true differential effect.) The model used to assess this interaction included a random slope for the indicator of unit (coded 0 for Ecology and 1 for Evolution) following Jaciw, Lin and Ma (2016); however, we also tested a model with a fixed slope which yielded an estimate of .176 standardized effect size units (t=2.25) that was statistically significant (p=.025).

This result is important because it confirms *within* biology the prediction of how *EU* works differentially based on content type that was motivated by the finding of positive impact in U.S. History and no impact in Biology (where Evolution and Ecology outcomes were considered together). We must interpret these results as exploratory, given that the analyses were not identified before the study. (However, the hypothesis for results by biology content area was made in advance of seeing the results presented here.)

Note that we also analyzed impacts separately by Ecology and Evolution units within biology using the benchmark impact model (but without repeated measures for individuals). The point estimate for the impact of *EU* on the Evolution unit was 1.68 scale score units (p=.499), based on 368 students with posttests for this outcome. The point estimate for the impact of *EU* on the Ecology unit was -2.05 scale score units (p=.333), based on 372 students with posttests for this outcome (Table 16). This means the impact is slightly positive for Evolution, and slightly negative for Ecology, and neither of these impacts is statistically significant; however, the difference in impact between them is marginally significant and favors Evolution.

These results should be addressed in full when considering the findings of the primary distinction between U.S. History and biology, and when evaluating the theory for observing a difference in impact between Ecology and Evolution. While the difference between these subdomains is marginally statistically significant and in the predicted direction, impact in either subdomain is not significantly different from zero. The theory should address not just why there is a difference in impact, but why impact is not significantly different from zero for either subdomain.

TABLE 16. RESULTS FOR BIOLOGY OUTCOME IMPACT ANALYSIS, BY UNIT

	Condition	Means	Standard deviations[a]	No. of posttest scores	No. of students	No. of teachers	Effect size	*p* value	Change in percentile ranking
Evolution Unit Outcome									
Unadjusted effect size[a]	Control	73.10	22.54	198	198	9	.09	.686	4%
	EU	75.08	22.03	170	170	9			
Adjusted effect size[b]	Control	75.37					.08	.499	3%
	EU	77.06							
Ecology Unit Outcome									
Unadjusted effect size[a]	Control	68.45	22.17	198	198	9	-.10	.604	-4%
	EU	66.30	21.58	174	174	9			
Adjusted effect size[b]	Control	68.45					-.09	.333	-4%
	EU	66.40							

[a] The unadjusted effect size is the regression-adjusted impact estimate in a model without covariates divided by the pooled standard deviation in outcomes.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Note. Full estimates corresponding to the impact models are reported in Appendix I.

Source. Empirical Education staff calculations

For completeness we also analyzed impacts separately for the World War II and Cold War enhanced units in U.S. History, using the benchmark impact model (but without repeated measures for individuals). The point estimate for the impact of *EU* on the World War II unit was 7.33 scale score units (p=.038), based on 227 students with posttests for this outcome. The point estimate for the impact of *EU* on the Cold War unit was 5.63 scale score units (p=.087), based on 227 students with posttests for this outcome (Table 17).

TABLE 17. RESULTS FOR U.S. HISTORY OUTCOME IMPACT ANALYSIS, BY UNIT

	Condition	Means	Standard deviations[a]	No. of posttest scores	No. of students	No. of teachers	Effect size	*p* value	Change in percentile ranking
World War II Unit Outcome									
Unadjusted effect size[a]	Control	53.05	24.59	122	122	6	.35	.172	14%
	EU	61.11	20.77	105	105	6			
Adjusted effect size[b]	Control	53.05					.32	.038	13%
	EU	60.38							

TABLE 17. RESULTS FOR U.S. HISTORY OUTCOME IMPACT ANALYSIS, BY UNIT

	Condition	Means	Standard deviations[a]	No. of posttest scores	No. of students	No. of teachers	Effect size	*p* value	Change in percentile ranking
Cold War Unit Outcome									
Unadjusted effect size[a]	Control	45.67	18.77	120	120	6	.31	.304	12%
	EU	51.48	18.94	107	107	6			
Adjusted effect size[b]	Control	45.67					.30	.087	12%
	EU	51.30							

[a] The unadjusted effect size is the regression-adjusted impact estimate in a model without covariates divided by the pooled standard deviation in outcomes.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Note. Full estimates corresponding to the impact models are reported in Appendix I.

Source. Empirical Education staff calculations

Treatment-Control Contrast

We examined whether there was a difference between *EU* and control classes in the use of the four SIM routines deemed central to implementation of *EU*.[8] (The results here build on the descriptives reported earlier for the *EU* condition.) Their use in the control conditions would indicate contamination and a reduced treatment-control contrast. We display the frequency of use in each condition for U.S. History, biology and both subjects combined in Figures 5 to 10, below. We observe strong contrasts, except for the Unit Organizer which is used in the control group but at a lower frequency than in *EU* classes. The teachers who responded that they used the Unit Organizer at least "Sometimes" in their control classes were all from one district with prior SIM exposure. Additionally, on each of the instructional logs, teachers were asked if they "intentionally or unintentionally use any of the Enhanced Units tools or strategies in your most recent control classes? (Tools include CORGI and the unit organizer, and strategies include Cause/Effect, Concept Comparison, and Question Exploration)." Teachers were further directed to "only answer 'Yes' if you had NOT learned about these tools/strategies prior to the Enhanced Units training. If you used these tools/strategies prior to your involvement with the Enhanced Units program, these are considered part of your 'Business As Usual' pedagogy, therefore you may answer 'No'." All teachers responded "No" to this question on each of the instructional logs across the study units. Therefore, we have no evidence of contamination based on the definition above.

"

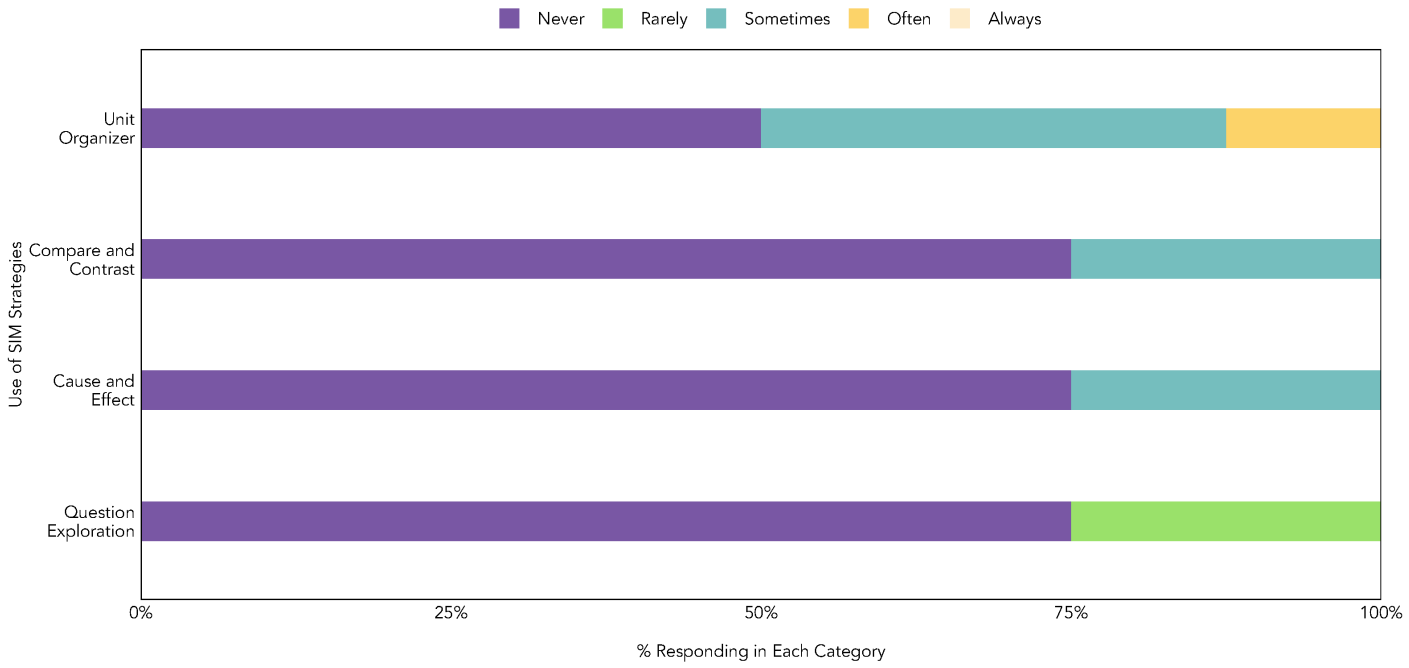


FIGURE 5. USE OF SIM PRACTICES, U.S. HISTORY TEACHERS IN CONTROL CLASSES ACROSS UNITS

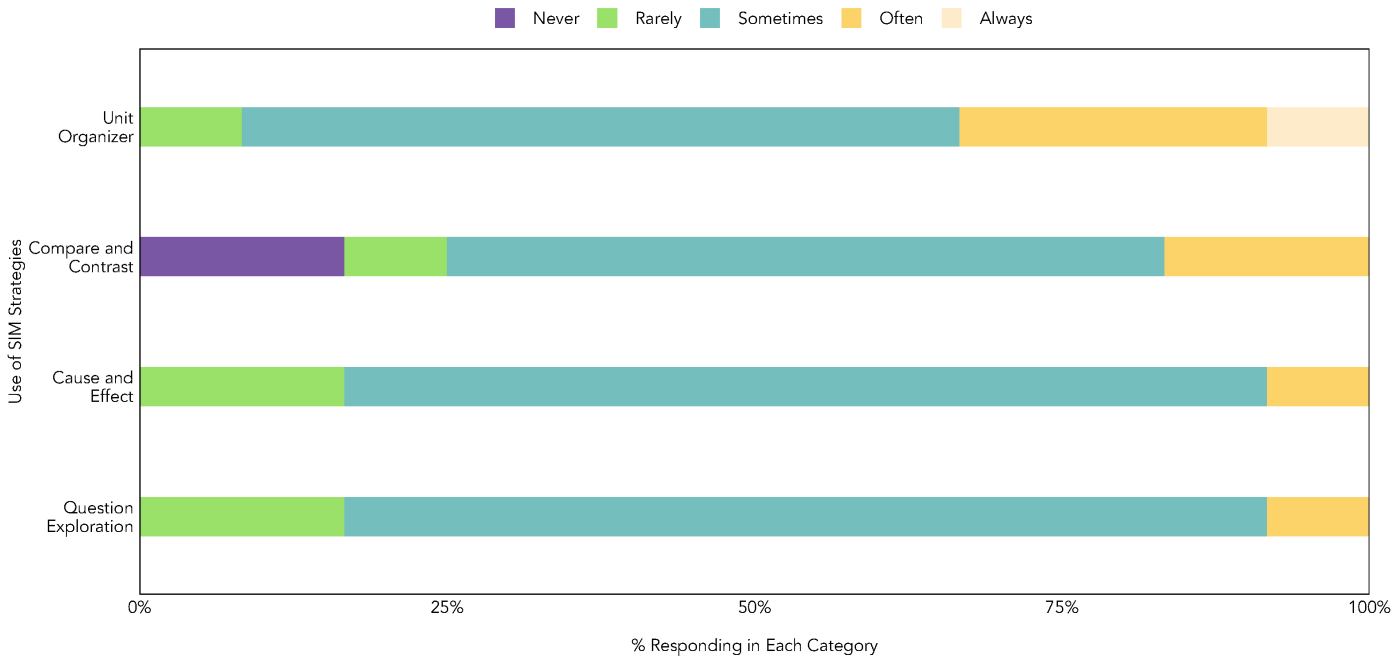


FIGURE 6. USE OF SIM PRACTICES, U.S. HISTORY TEACHERS IN TREATMENT CLASSES ACROSS UNITS

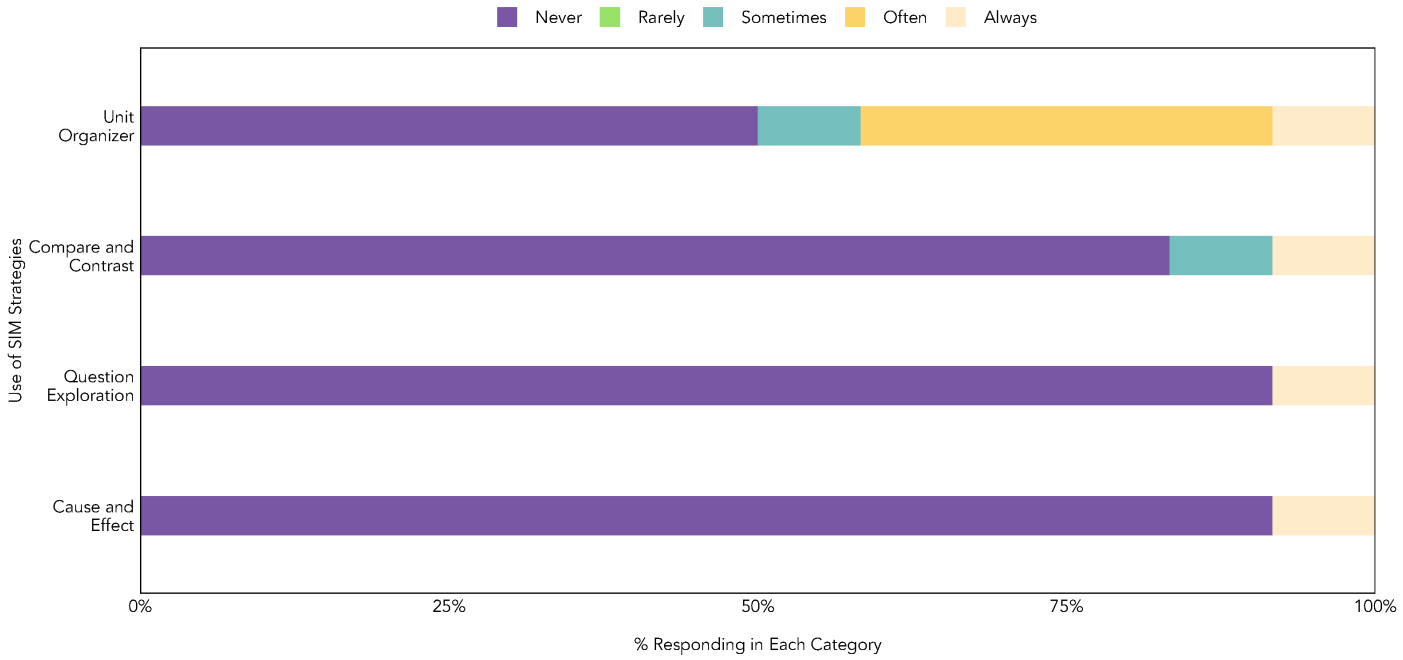


FIGURE 7. USE OF SIM PRACTICES, BIOLOGY TEACHERS IN CONTROL CLASSES ACROSS UNITS

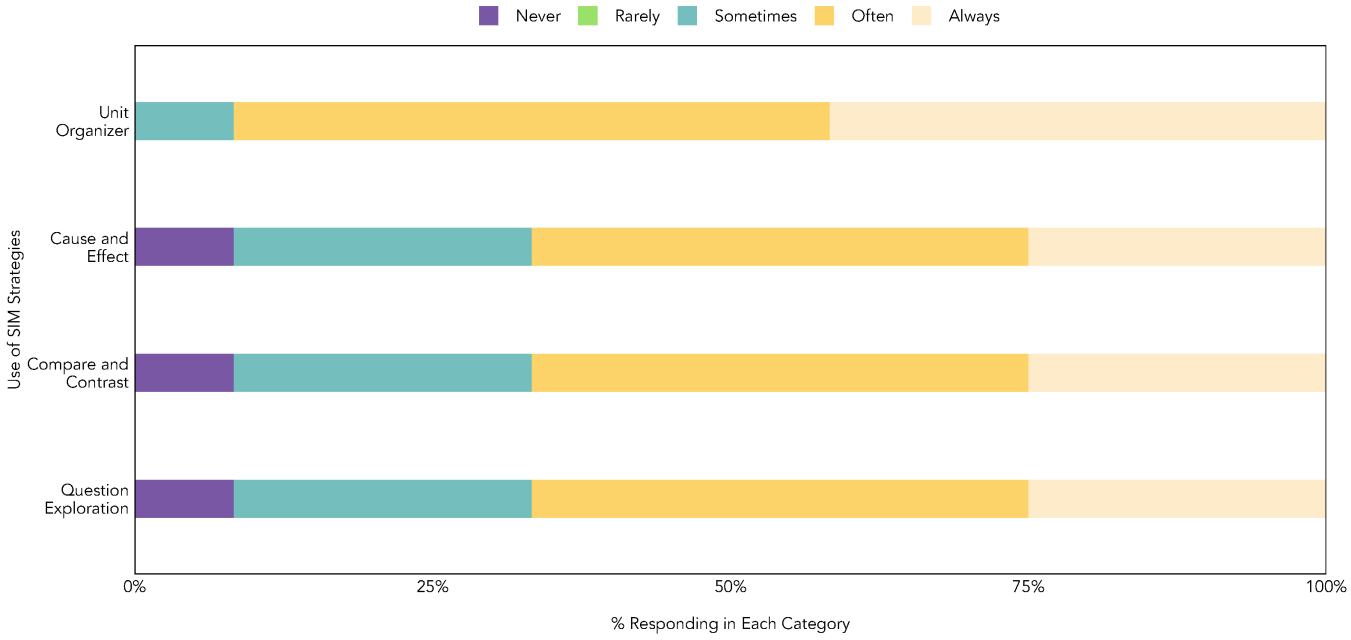


FIGURE 8. USE OF SIM PRACTICES, BIOLOGY TEACHERS IN TREATMENT CLASSES ACROSS UNITS

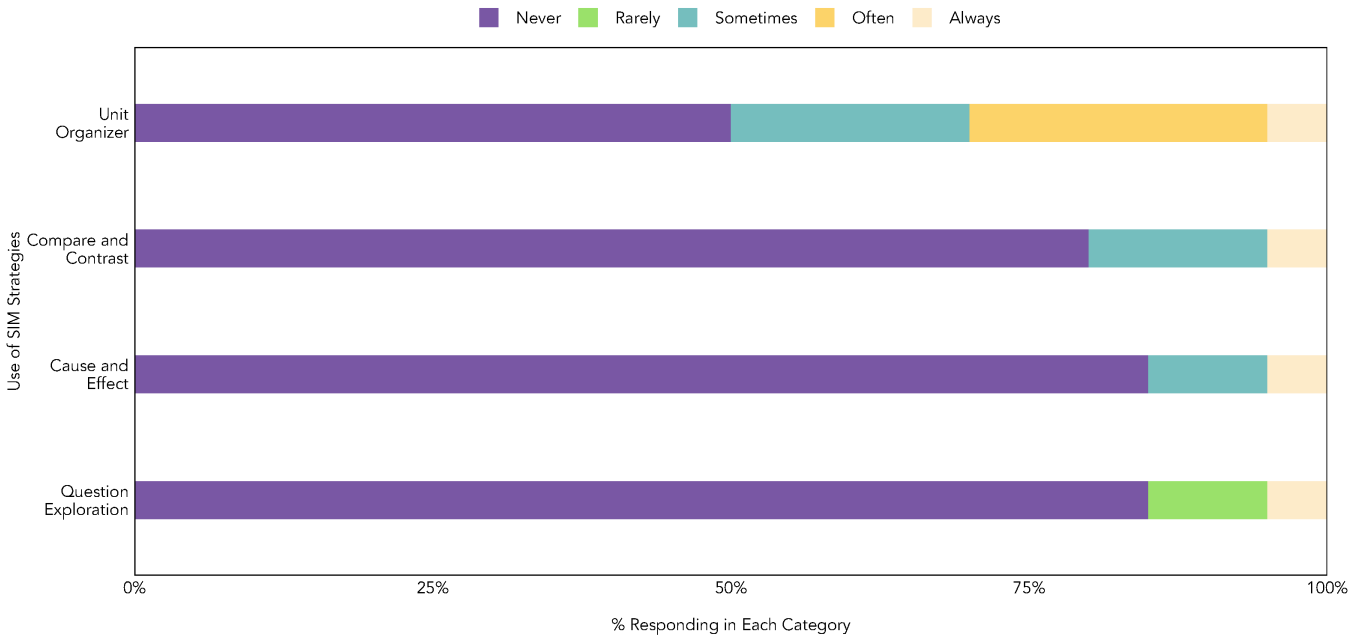


FIGURE 9. USE OF SIM PRACTICES, ALL TEACHERS IN CONTROL CLASSES ACROSS UNITS

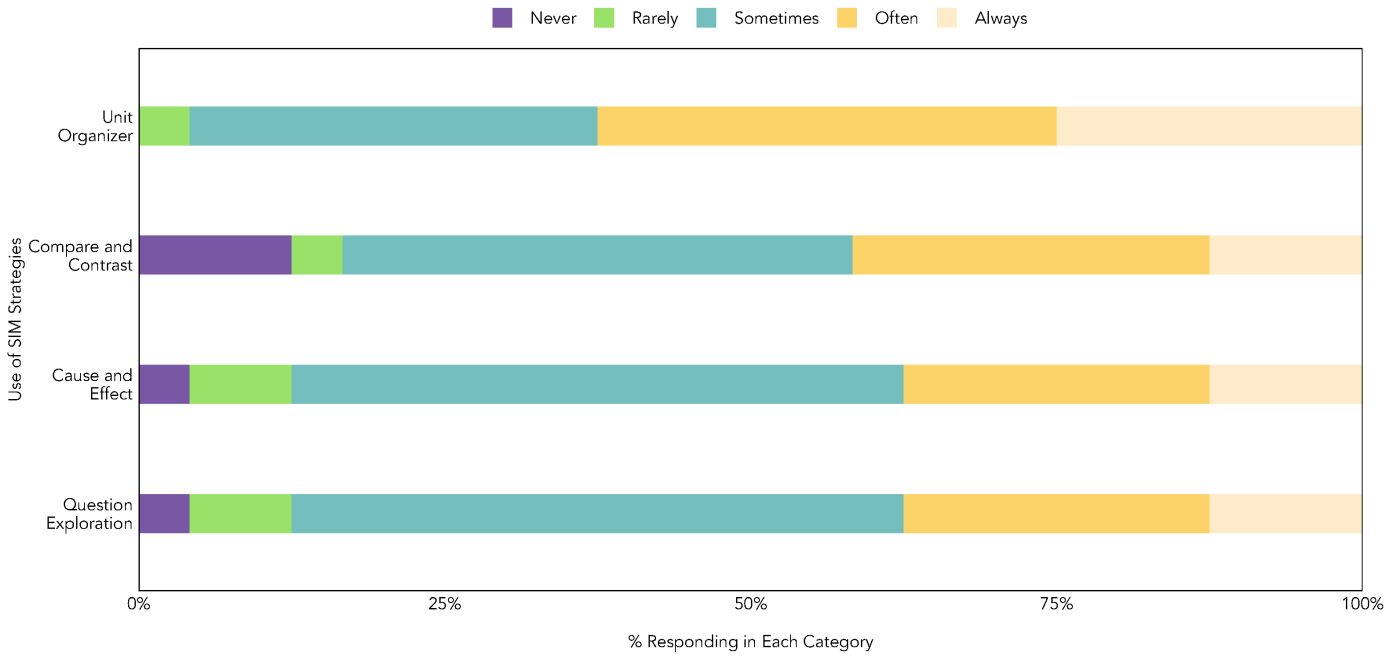


FIGURE 10. USE OF SIM PRACTICES, ALL TEACHERS IN TREATMENT CLASSES ACROSS UNITS

For the combined sample (i.e., comparing all treatment classes to all control classes across subject areas), the differences in the use of SIM instructional practices across conditions was statistically significant: Use of Unit Organizers (p = .017), Compare and Contrast (p = .009), Cause and Effect (p = .002), Question Exploration (p = .001).

Connections between Fidelity of Implementation and Impact

We investigated the relationship between FOI and impact. To do this we examined the correlation across randomized blocks between the *EU*-control difference in average student performance and block-level FOI (i.e., fidelity scores by component for the treatment class within each block). We discussed the rationale for this approach earlier under Approach to Analysis. Specifically, in these results, we look for a positive trend in the correlation between the two variables, recognizing that differences in the *EU*-control achievement outcomes are some combination of remaining class-level sampling variation and variation in impact of *EU* across randomized blocks. (The results are explained in greater detail in Appendix K.)

Figure 11 through Figure 16 exhibit correlations between: (1) block-specific regression-adjusted estimates of the difference between *EU* and control classes in outcomes using the test metric (on the X-axis), and (2) average FOI measure (on Y-axis) for each of six FOI measures listed here.

- FOI1: Number of minutes of *EU* coaching received over the study. [The threshold set by SRI/CAST for meeting fidelity was 8 hours (480 minutes)].
- FOI2: Weighted average of points earned for using each of the four SIM routines. (Minimum zero, maximum 18.)
- FOI3: Weighted average of points earned for combining each of the four SIM routines with CORGI and co-construction (working collaboratively with students). (Minimum zero, maximum 12.)
- FOI4: Teacher reported average of helpfulness of *EU*. (On a Likert scale, with minimum 1, maximum 5).
- FOI5: Student reported average of how *EU* improves understanding of content. (On a Likert scale with minimum 1, maximum 5.)
- FOI6: Student reported average of how CORGI helps students collaborate. (On Likert scale, with minimum 1, and maximum 5.)

The correlations for the six measures of FOI and the block-specific estimate of impact are: .63 (p=.02), .18 (p=.54), .19 (p=.52), .012 (p=.97), -.12 (p=.72) and -.19 (p=.55). Results from U.S. History classes are in yellow, and biology are in purple. (One teacher chose to not participate in *EU* following the practice unit. We included implementation results where available for the two randomized blocks for that teacher. For FOI2 and FOI3 fidelity scores take value 0; for FOI4 – FOI6 there are no fidelity scores for this teacher.)

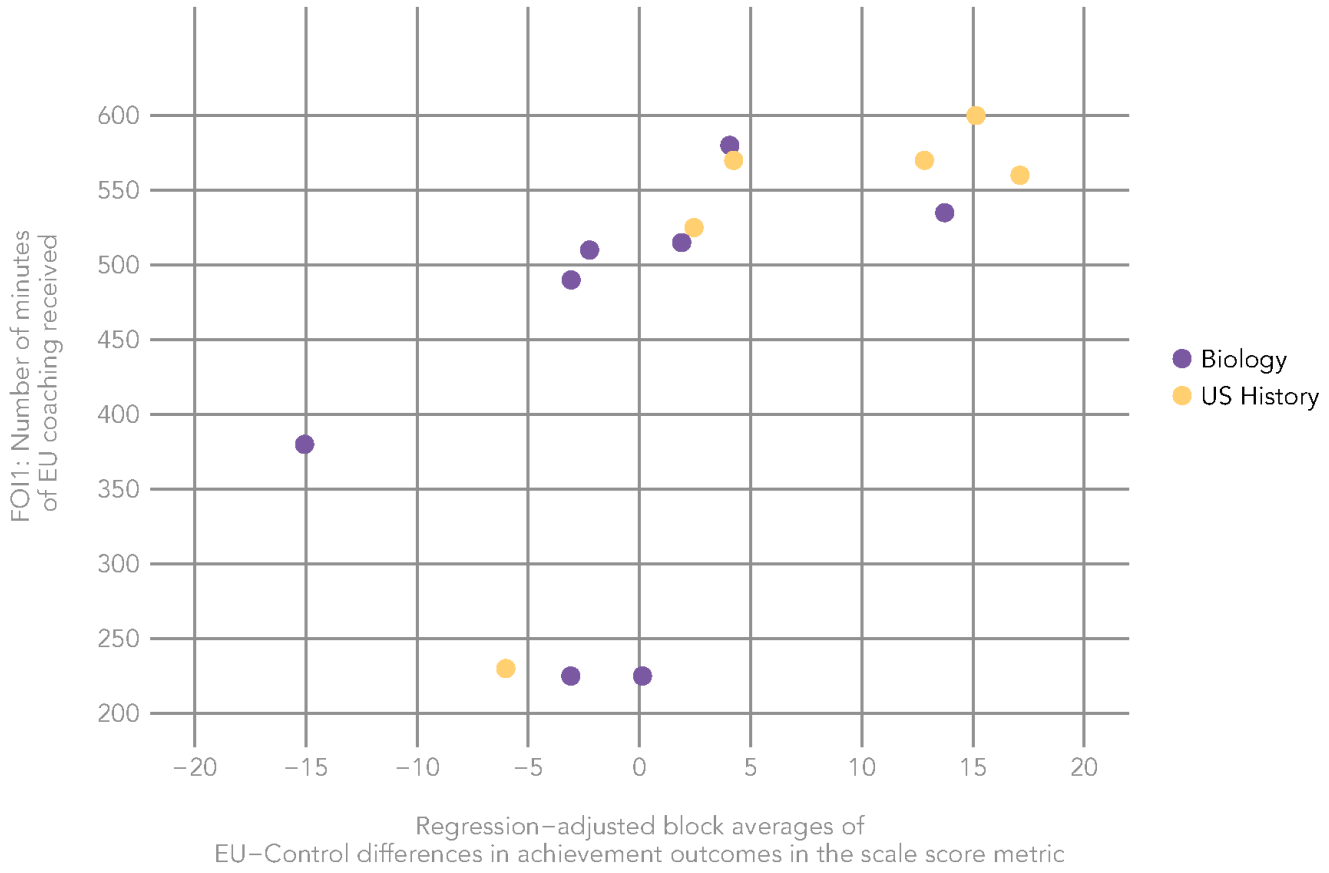


FIGURE 11. CORRELATION BETWEEN FOI1 (TEACHERS RECEIVE COACHING) AND IMPACT ON ACHIEVEMENT ACROSS RANDOMIZED BLOCKS

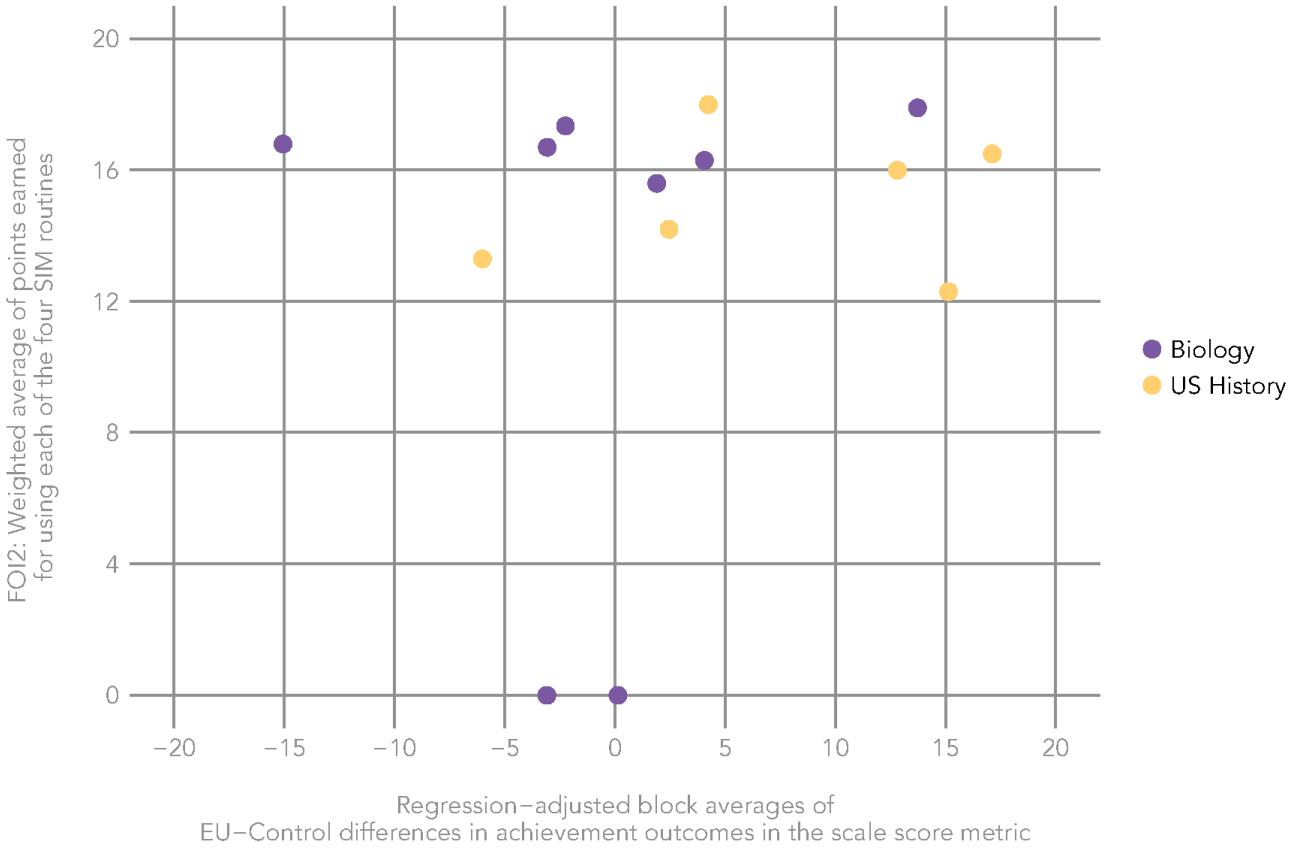


FIGURE 12. CORRELATION BETWEEN FOI2 (ADHERENCE) AND IMPACT ON ACHIEVEMENT ACROSS RANDOMIZED BLOCKS

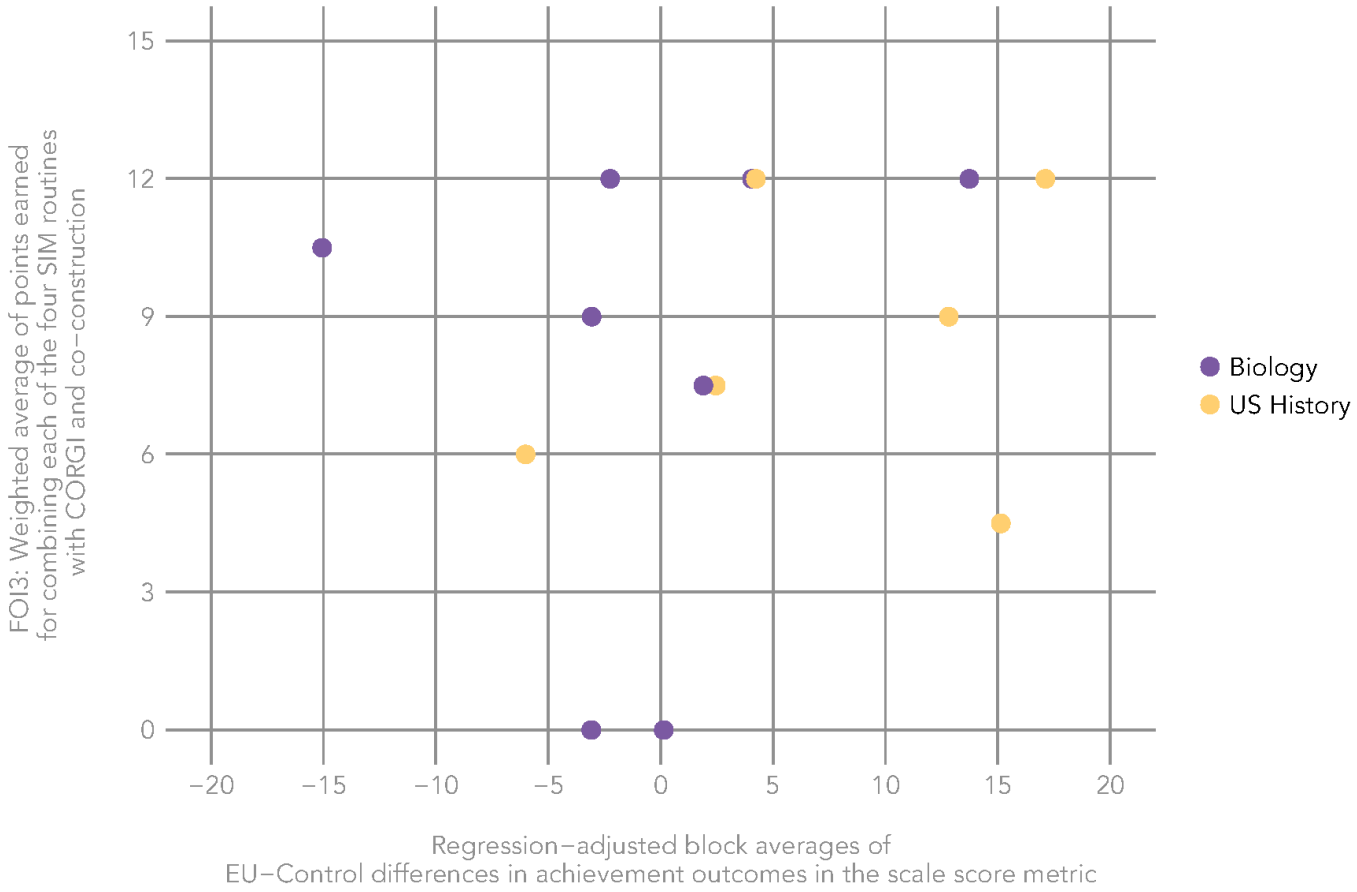


FIGURE 13. CORRELATION BETWEEN FOI3 (QUALITY) AND IMPACT ON ACHIEVEMENT ACROSS RANDOMIZED BLOCKS

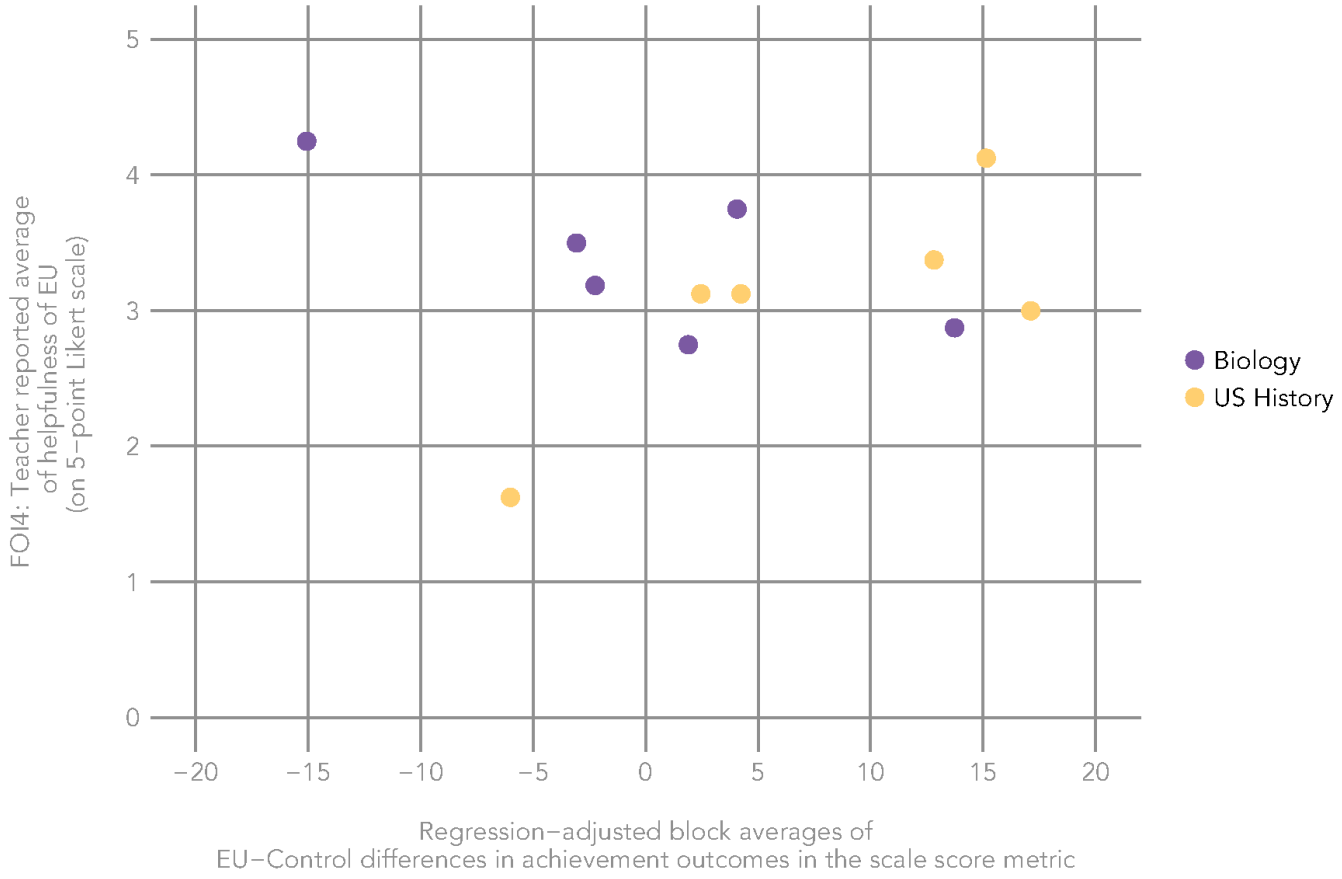


FIGURE 14. CORRELATION BETWEEN FOI4 (USEFULNESS OF *EU*) AND IMPACT ON ACHIEVEMENT ACROSS RANDOMIZED BLOCKS

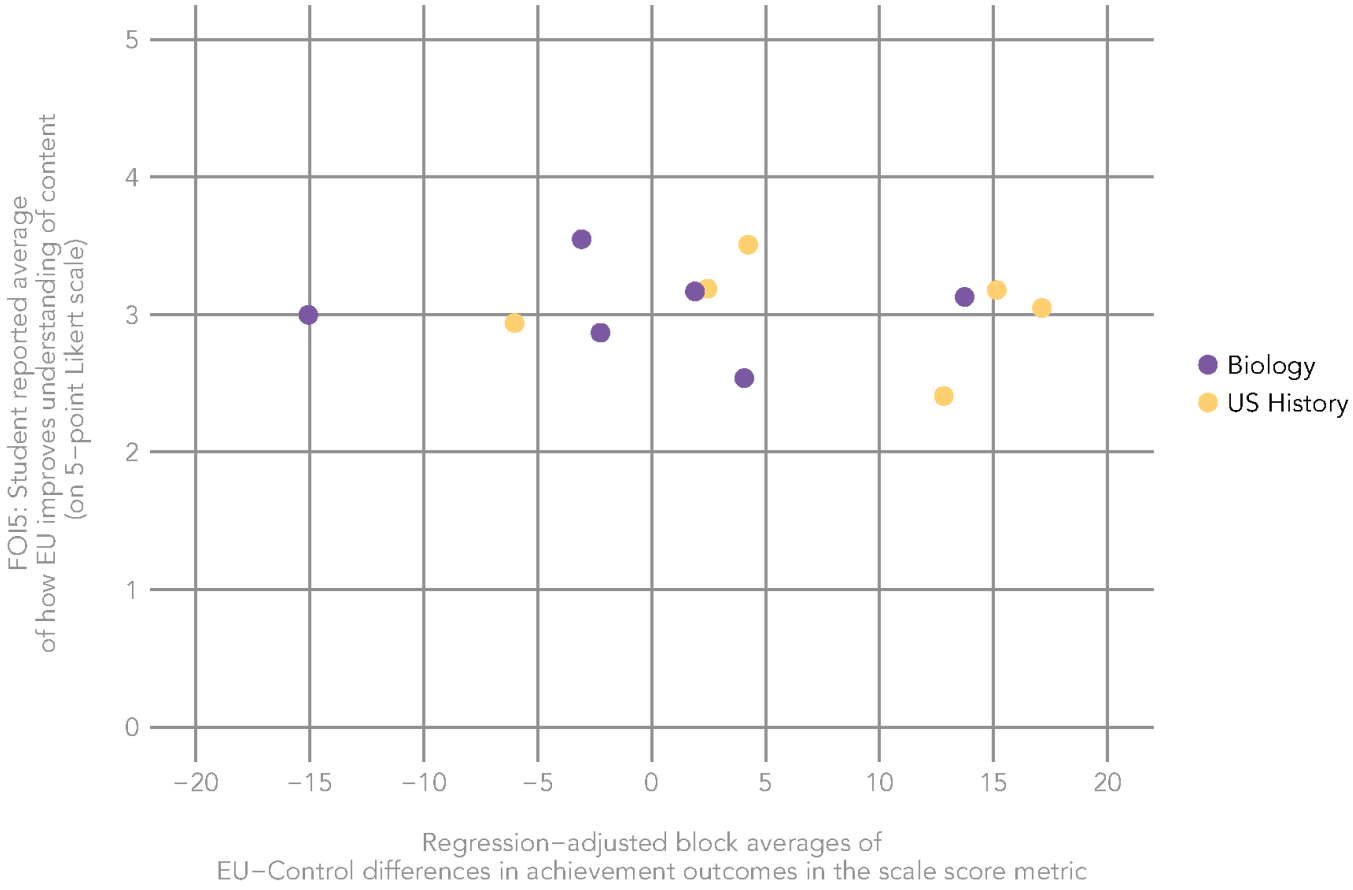


FIGURE 15. CORRELATION BETWEEN FOI5 (STUDENT UNDERSTANDING) AND IMPACT ON ACHIEVEMENT ACROSS RANDOMIZED BLOCKS

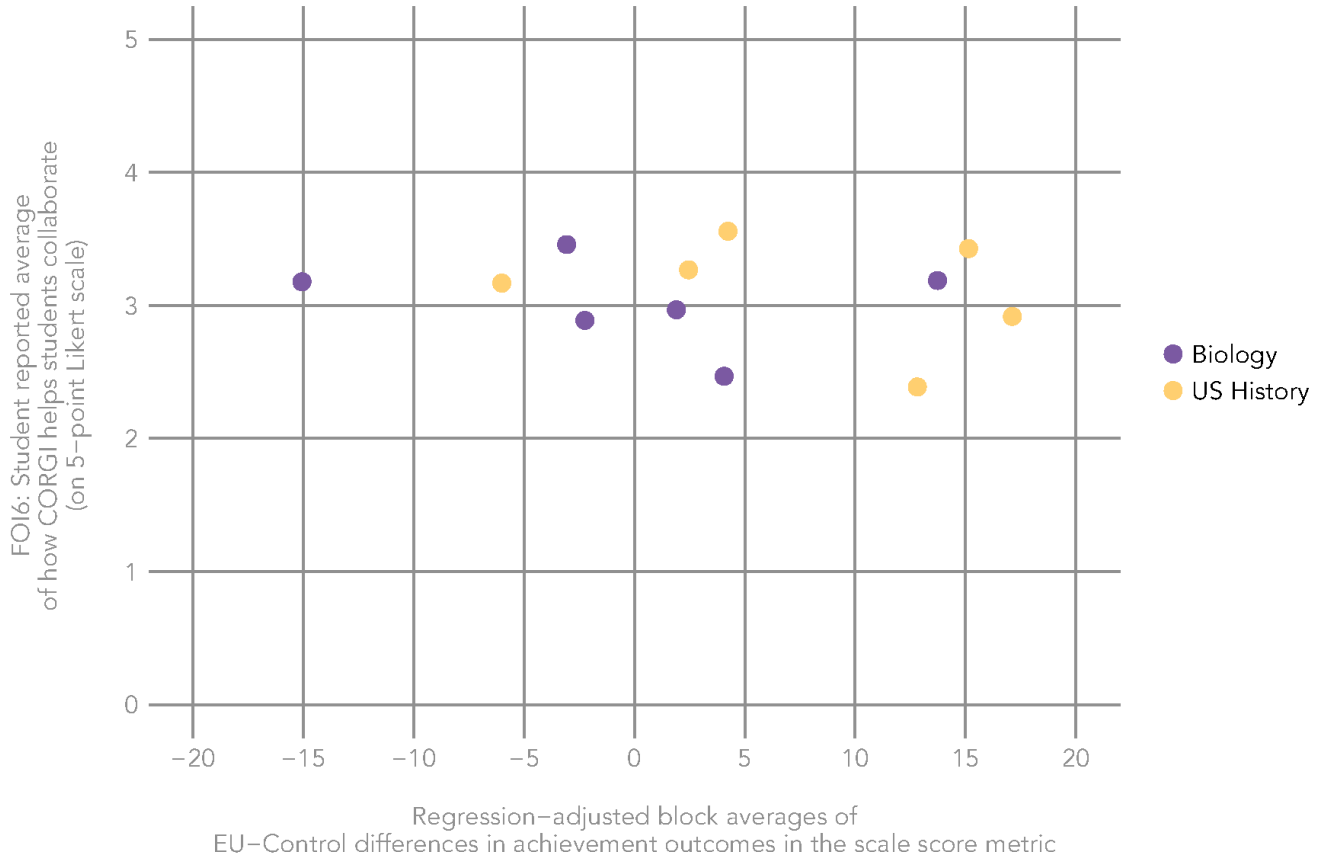


FIGURE 16. CORRELATION BETWEEN FOI6 (STUDENT COLLABORATION) AND IMPACT ON ACHIEVEMENT ACROSS RANDOMIZED BLOCKS

Impact on Mediators (for the Combined Sample and by U.S. History and Biology)

Table 18 below reports the sample mean and median (across blocks) in the difference between *EU* and control in frequency of use of each of the 17 mediating practices identified by SRI. Figure 17 and Figure 18 display the same information in graphical form. (Given the small samples involved, any inferential test will be underpowered. We observe very little difference between *EU* and control in their practices, with values very close to zero compared to maximum possible differences ranging between -4 and +4.)

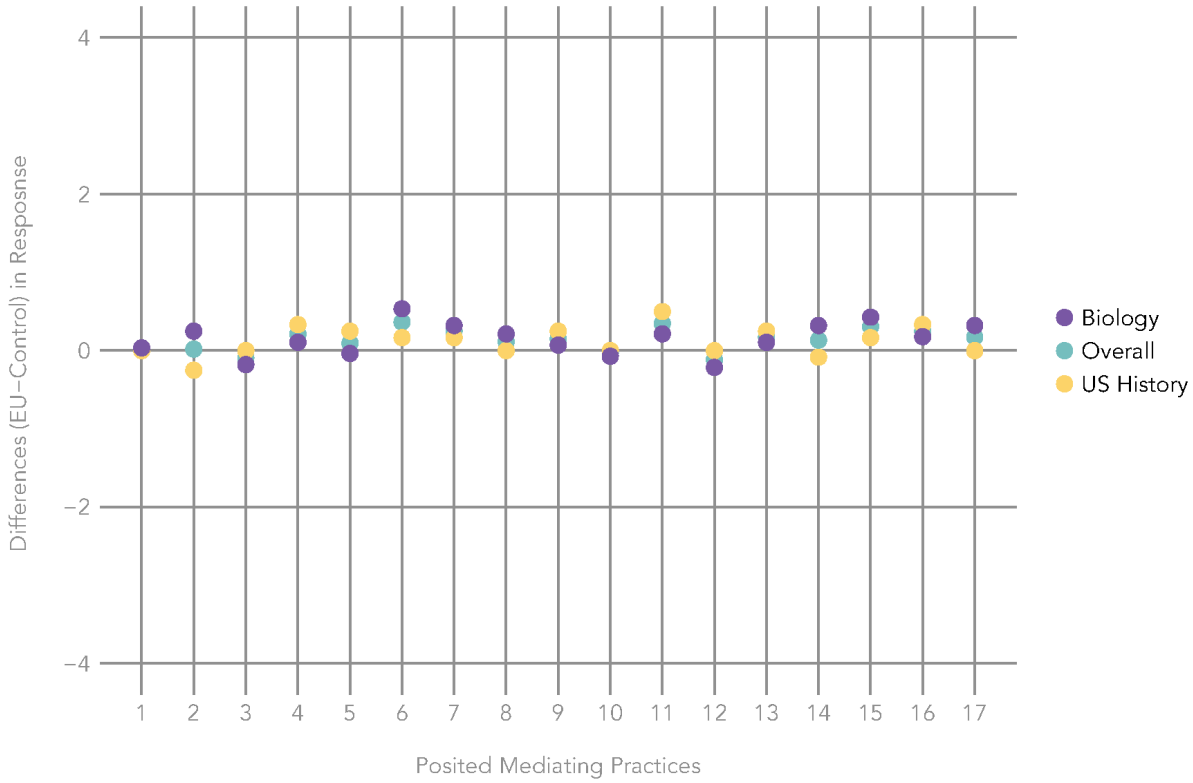


FIGURE 17. AVERAGE ACROSS RANDOMIZED BLOCKS IN DIFFERENCES (*EU* – CONTROL) IN ORDINAL RESPONSES TO FREQUENCY OF USE OF EACH OF 17 POTENTIAL MEDIATING PRACTICES

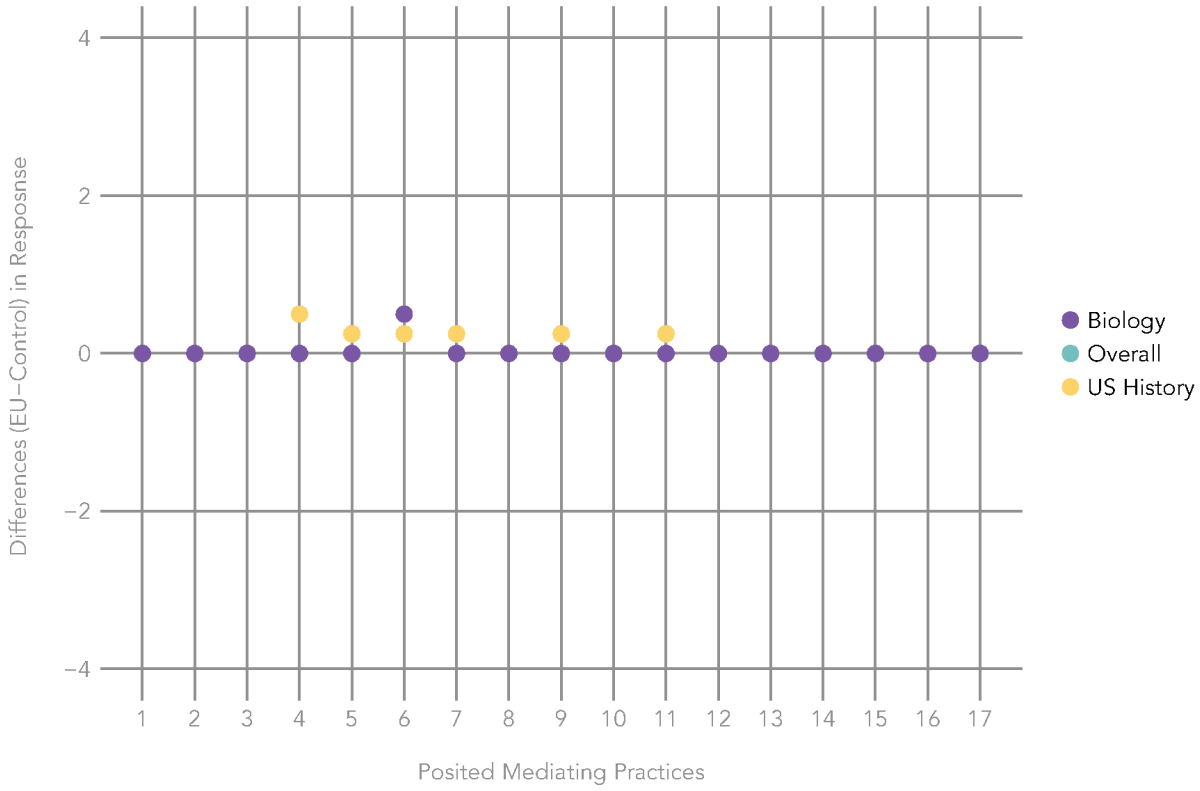


FIGURE 18. MEDIAN ACROSS RANDOMIZED BLOCKS IN DIFFERENCES (*EU* – CONTROL) IN ORDINAL RESPONSES TO FREQUENCY OF USE OF EACH OF 17 POTENTIAL MEDIATING PRACTICES

TABLE 18. MEAN AND MEDIAN DIFFERENCES (*EU* – CONTROL) IN ORDINAL RESPONSES TO FREQUENCY OF USE OF EACH OF 17 POTENTIAL MEDIATING PRACTICES

		Mean			Median		
	Description	Overall N = 13	U.S. History N = 6	Biology N = 7	Overall N = 13	U.S. History N = 6	Biology N = 7
1	**Explicit instruction**	.02 (*p*=.625)	.00	.04	.00	.00	.00
2	**Reteach to a few students**	.02 (*p*=.875)	-.25	.25	.00	.00	.00
3	**Identifying similarities/differences (non-SIM)**	-.10 (*p*=.322)	.00	-.18	.00	.00	.00
4	**Explicit strategy for asking clarifying questions (non-SIM)**	.21 (*p*=.781)	.33	.11	.00	.50	.00
5	**Explicit summarizing strategy (non-SIM)**	.10 (*p*=1.00)	.25	-.04	.00	.25	.00
6	**Explicit paraphrasing strategy (non-SIM)**	.37 (*p*=.424)	.17	.54	.50	.25	.50
7	**Explicit vocabulary strategy (non-SIM)**	.25 (*p*=.359)	.17	.32	.00	.25	.00
8	**Graphic organizer (non-SIM)**	.12 (*p*=1.00)	.00	.21	.00	.00	.00
9	**Note-taking technique**	.15 (*p*=.625)	.25	.07	.00	.25	.00
10	**Mnemonic device for remembering information**	-.04 (*p*=.375)	.00	-.07	.00	.00	.00
11	**Rehearsing information aloud**	.35 (*p*=.156)	.50	.21	.00	.25	.00
12	**Teacher laptop or Chromebook**	-.12 (*p*=.375)	.00	-.21	.00	.00	.00
13	**Student laptop or Chromebook**	.17 (*p*=.906)	.25	.11	.00	.00	.00
14	**Student tablet**	.13 (*p*=.750)	-.08	.32	.00	.00	.00
15	**Student collaboration on group and partner assignments**	.31 (*p*=.250)	.17	.43	.00	.00	.00
16	**Teaching higher-order course content**	.25 (*p*=.688)	.33	.18	.00	.00	.00
17	**Support for learners with different abilities**	.17 (*p*=.563)	.00	.32	.00	.00	.00

Note. Teacher responses were on an ordinal scale: Never, Seldom, Sometimes, Often, and Always; *p* values for differences in means are based on Wilcoxon Signed Rank Test. The overall means were calculated across blocks, weighting blocks equally.

Source. Empirical Education staff calculations

Discussion

We found a positive impact of *Enhanced Units* (*EU*) on student achievement in U.S. History, but not on biology or on outcomes combined across the two domains. The impact on U.S. History was .32 standardized effect size units.

We also examined whether the impact of the program on the combined sample (U.S. History and biology) varied depending on characteristics of teachers and students. We found no difference in the impact depending on a teacher's facility with technology reported at baseline. We did, however, find a positive differential effect favoring students with disabilities. This is an encouraging result given that a primary goal of the grant was to support students with disabilities or other learning challenges.

Were conditions to support seeing an impact present? We found that certain conditions supporting impact were satisfied: the treatment-control contrast was strong (there was limited spillover in the use of *EU* by teachers in their control classes), and there was a moderate positive correlation between minutes of *EU* coaching and block-specific impacts for the combined sample. However, other results made it harder to explain the mechanism behind the observed impact. Implementation did not reach threshold levels of fidelity, system-wide as developers expected. There was little correlation between other measures of fidelity at the block level (e.g. levels of reported helpfulness) and block-specific impact (although this analysis had limited power). Furthermore, though the study was too underpowered to conduct a formal mediation analysis, we saw little difference between *EU* and control in teacher practices identified as potential mediators of impact on achievement (Table 18).

A further question is why we observed a positive impact for U.S. History but not Biology. A theory developed by the program developers is that *EU* works especially well with content that progresses in a sequential and linear way. We were able to test this hypothesis by examining whether, within biology, we would see a greater impact for the unit on evolution than for the unit on ecology, as the content and routines in the former were structured in a more-sequenced way. Impact in ecology was greater by .171 standard deviation units (t=2.00, p=.063), confirming the hypothesis with moderate confidence.

Given the differences in impact observed both between subject areas and across units within a subject area, and no clear impacts on posited mediators, there is reason to continue to explore the contexts and conditions for the observed effects. There is also need to further understand how implementation, as defined in this work, relates to impact or lack thereof. This additional work is recommended through a follow-on project.

Where can program improvement efforts be focused? If *EU* works better with logically sequenced material, as the results seem to suggest, then the obvious place to focus improvement is with unstructured material. That is, we should seek program development to support impact for material with wide-ranging structures, and possibly an introductory element that systematically links the content to previously learned content.

Another area where teacher comments in interviews suggest improvement efforts is the usage of the routines in conjunction with CORGI. When asked about the extent to which they did or did not find CORGI to be useful, 9 out of 12 teachers' answers contained statements suggesting that they felt CORGI hindered the usefulness of the routines. In terms of CORGI's logistics, teachers cited the visual interface, usability, and difficulties navigating the Google Drive environment to save, share, and access files as obstacles. These are operational issues that are likely best addressed by the user design team. But in terms of CORGI's acute effects on students' experiences, reasons cited for *not* being useful

included its inability to accommodate or engage students of all abilities, students who struggled with typing, and students who preferred doing the routines on paper.

Moreover, it is useful to consider how teachers responded to being asked if they would recommend *EU* to other teachers. The one teacher who answered "no," cited inadequate planning time and lack of student buy-in. Eleven out of twelve teachers interviewed said "yes," but all of them with various conditions or suggestions. These included: (1) teachers receiving adequate training in the routines, (2) classes having access to user-friendly computer devices, (3) teachers having more discretion over pairing routines with topics taught, (4) routines being done on paper instead of CORGI, and (5) *EU* being used for Advanced Placement, honors, or upper-level classes.

Future improvements to *EU* should focus on answering the question: "What is/are the best way(s) for teachers to present SIM routines to their students, particularly for students with learning challenges through SIM intervention?" While the first and second points listed above (training and devices) can be addressed through adequate program implementation, the remaining three points invite further exploration by the program developer. Assuming that teachers have greater discretion over pairing routines with topics taught, program development should investigate how the routines can be applied to a greater range of topics. In regards to the routines being done on paper instead of CORGI, program development should consider how introducing devices to the routines potentially presents steeper learning curves and difficulty with buy-in for teachers and students alike. Perhaps, it would be wise to consider a teacher's assertion, "I almost feel like sometimes we try to use too much technology when simple is sometimes better." Finally, with respect to the suggestion that *EU* be used in upper-level classes, program development should consider how this should be balanced with the priority of the grant: improving academic outcomes for students with learning disabilities through SIM interventions.

References

Deshler, D. D., & Schumaker, J. B. (Eds.). (2006). *Teaching adolescents with disabilities: Accessing the general education curriculum*. Thousand Oaks, CA: Corwin Press.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology, 76*(4), 569-582. With Permission. Retrieved from http://u.osu.edu/hoy.17/research/instruments/#Long

Jaciw, A., Lin, L., & Ma, B. (2016, October). An empirical study of design parameters for assessing differential impacts for students in group randomized trials. *Evaluation Review 40*(5), 410-443. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/0193841X16659600

Penuel, W. R., & Martin, C. (2015, April). *Design-Based Implementation Research as a strategy for expanding opportunity to learn in school districts.* Invited presentation to the Research Conference of the National Council of Teachers of Mathematics, Boston, MA.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

SAS Institute Inc. (2006). *SAS/STAT Software: Changes and Enhancements.* (Through release 9.1). Cary, NC: Author.

Schmidt, D. A., Baran, E., Thompson, A. D., Mishra, P., Koehler, M. J., Shin, T.S. (2009). Technological Pedagogical Content Knowledge (TPACK): The Development and Validation of an Assessment Instrument for Preservice Teachers. *Journal of Research on Technology in Education, 42*(2), 123-149. Retrieved October 3, 2018 from https://files.eric.ed.gov/fulltext/EJ868626.pdf

Schochet, P. Z. (2009). *Do Typical RCTs of Education Interventions Have Sufficient Statistical Power for Linking Impacts on Teacher Practice and Student Achievement Outcomes?* (NCEE 2009-4065). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Tschannen-Moran, M., & Hoy, A. W. (2001, February). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*(7), 783-805. Retrieved from https://pdfs.semanticscholar.org/42e3/af5061de43d25b85144964ab7a3e44a7549f.pdf

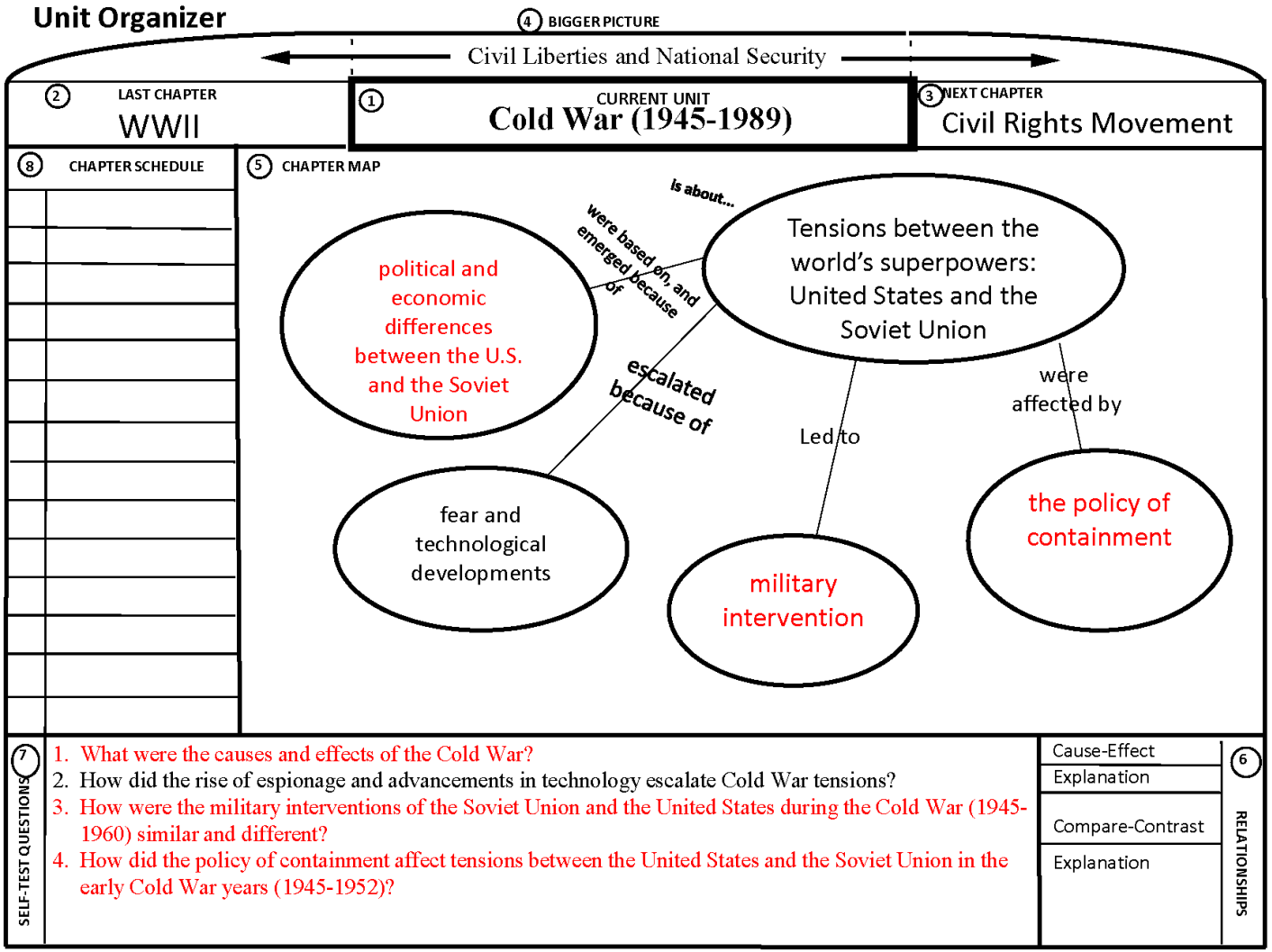
What Works Clearinghouse. (2017). Standards handbook (Version 4.0). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

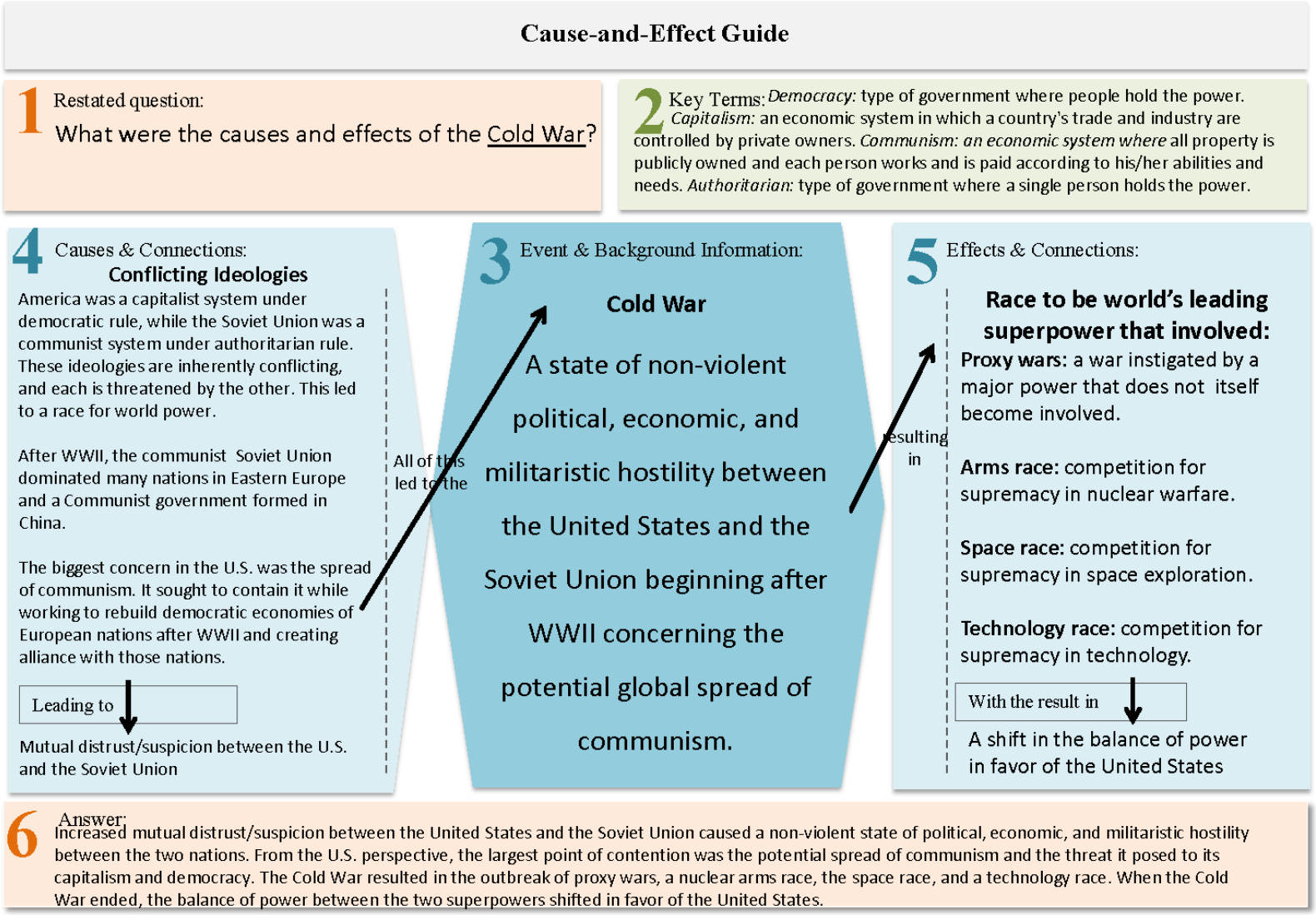
Appendix A. Examples of the Content Enhancement Routines and Definitions Used in Report

Cold War Enhanced Unit

Redesigning Secondary Courses to Improve Outcomes for Adolescents with Disabilities and other Underperforming Adolescents
Department of Education I3 (Investing in innovation) PR/Award (U411C14000)

Red print indicates routines to be used in pilot study.





Comparison Table

C Communicate targeted concepts
O Obtain the Overall Concept
M Make lists of known characteristics
P Pin down Like Characteristics
A Assemble Like Categories
R Record Unlike Characteristics
I Identify Unlike Categories
N Nail down a summary
G Go beyond the basics

② Overall Concept

How were the military interventions of the Soviet Union and the United States during the early Cold War (1945-1960) similar and different?

① Concept

Military interventions of the Soviet Union

③ Characteristics

- Installed satellite nations (those under Soviet control) to prevent future invasions and influence from the U.S.
- Blockaded West Berlin from Western nations
- Formed Warsaw Pact with seven communist Eastern European nations.
- Soviet Union exploded atomic bomb
- Built nuclear arsenal
- Soviets sent troops from Communist N. Korea into South Korea
- Soviet Union gave communist China ammunition and guns surrendered by the Japanese to fight against the nationalists
- Suppressed democratic revolt in communist Hungary

① Concept

Military interventions of the United States

③ Characteristics

- Adopted policy of containment: measures to prevent expansion of communism to other countries and curb Soviet influence.
- Berlin Airlift—flew food and supplies into West Berlin
- Joined NATO—a military alliance created among democratic nations during peacetime following WWII.
- U.S. exploded H-bomb
- Built nuclear arsenal
- U.S. sent troops to South Korea to stop the communist invasion
- U.S. gave communist opposition in China $2 billion worth of military equipment and supplies during its Civil War
- U.S. did not aid in the revolt in Hungary

⑨ Extensions

How did the Cold War end any hope the United States may have had to returning to isolationism?

How has the Cold War impacted the United States' relationship with North Korea and China?

④ Like Characteristics

- Sought to curb influence of opposing nation around the world
- Soviets isolated West Berlin, but U.S. airlifted food and supplies
- U.S. formed a military alliance with western European nations—NATO; Soviet Union formed the Warsaw Pact with eastern European nations.
- Showed threat by building nuclear arsenal
- Sent military in to Korea
- Aided opposing sides in Chinese Civil War

⑤ Like Categories

- Tried to curb other's influence
- Tried to control West Berlin
- Formed different alliances
- Built nuclear arsenal
- Sent troops to Korea
- Took opposing sides in China

⑥ Unlike Characteristics

- Soviets exploded A-bomb; U.S. exploded H-bomb
- Soviets suppressed a democratic revolt in communist Hungary while the U.S. did nothing.

⑦ Unlike Categories

- Exploded different kinds of bombs
- Responded differently to the revolt in Hungary

⑧ Summary

The military interventions of the Soviet Union and the United States were similar and different during the early Cold War (1945-1960). They both sought to curb influence of the other, tried to maintain control over West Berlin, formed alliances, developed bombs and built an arsenal, sent troops to Korea and took sides in China. They differed in the types of nuclear weapons they had, and in how they responded to the uprising in Hungary.

Question Exploration Guide

1 What is the Critical Question?

How did the policy of containment affect tensions between the U.S. and the Soviet Union in the early Cold War years (1945-1952)?

2 What are Key Terms and Explanations?

• Iron Curtain	• the notional barrier separating the Soviet bloc and the West during the Cold War
• Potsdam Conference	• meeting between Stalin, Churchill, and Truman in 1945 to negotiate terms for the end of World War II
• Containment	• the action or policy of preventing the expansion of a hostile country or influence

3 What are Supporting Question/Answers?

When did tensions arise?	Tensions rose immediately following the end of WWII through the early-1950s between U.S. and Soviet Union.
Why was it tense?	Both nations had emerged as superpowers that could greatly influence world events, yet their ideologies conflicted.
How were they different?	The U.S. was a capitalistic, democratic nation while the Soviet Union was communist under dictatorial rule.
What political policies affected these tensions?	The superpowers each tried using the United Nations (a peacekeeping body) as a platform to exercise influence over others. At Potsdam, the Soviet Union gained new territories where Stalin installed communist governments (buffers). When Truman became president, he adopted a policy of containment, preventing the expansion of communism to other countries. It was then that an "iron curtain" fell across Europe, often seen as the start of the Cold War.
Did Truman do anything else?	Truman asked Congress for economic and military aid to Greece and Turkey, and more money to protect the rest of Western Europe from Communist influence by strengthening the democracies.
What did Stalin do in response?	Stalin blocked Berlin, isolating it from Western Europe but U.S. airlifted supplies to the people living in West Berlin.
How did the U.S. respond?	The U.S. helped form the North Atlantic Treaty Organization (NATO), which pledged support to allies in Western European and North American countries.

4 What is the Main Idea answer to the critical question?

The U.S. policy of containment increased tensions by trying to contain communism, while the Soviet Union was trying to expand it. In addition, the U.S. spent billions of dollars to support democracies around the world and by forming alliances with other democratic nations. The Soviet Union meanwhile was building its own alliances with communist countries.

5 How can we use the Main Idea?

How do the political and military policies of the United States affect current relationships with different countries around the world?

6 Is there and Overall Idea? Is there a real-world use?

What is the current state of the relationship between the United States and Russia?

The program development team recognizes that over the years, the terms "strategy" and "routine" have acquired various definitions. Based on original research and development at the University of Kansas Center for Research on learning, however, these terms were used in specific ways that characterized the settings, instructors, target students, goals, type of instruction, and instructional supports for each. For the purpose of clarity in this reporting, the following definitions, characterizations, and distinctions are made.

Content Enhancement Routines, referred to in this report as routines, were designed for use in general education classrooms taught by expert content teachers whose classes contained a wide range of student achievement and abilities. The routines help teachers think about student learning needs and styles, select important information and ways of thinking in a course, develop instructional supports, and plan for collaborative discussion with all students. Groups of routines have been developed to help teachers plan and organize instruction, explain details from texts, understand critical concepts, and engage students in higher order thinking and reasoning. The latter were the focus of this study. The goal was to promote mastery of critical learning across different subjects and content areas based on standards. The routines in this study focused on higher order reasoning associated with exploring and understanding critical concepts or main ideas, causes and effects, comparisons, or argumentation with consideration given to decision-making, problem solving and other routines. All routines include explicit instructional prompts such as advance and post organizers, interactive graphic organizer development, thinking steps within the organizer that guide reasoning, and collaborative discussion in partnership among all students.

Learning Strategies, referred to in this report as *strategies* or *LSs*, were designed to help students become independent learners by learning and using *basic learning skills* needed to complete a variety of academic tasks. LSs were taught by trained teachers in special classes with smaller numbers of students. A learning strategy has been defined as an individual's approach to a task. For example, if students have difficulty recognizing multi-syllabic words in a science text, a teacher could teach the student a strategy for quickly breaking the unknown word into word parts (i.e., prefix, suffix, stem) and then blending and pronouncing the whole word. Other LSs help students identify words from text, write sentences, paraphrase what was read and self-question text. LS instruction involves teaching strategy steps to mastery using a mnemonic device that helps students recall the steps, verbal explanations by teachers about the steps of a strategy, modeling of its use, feedback on student performance, controlled practice, and generalization.

Appendix B. Detailed Description of DBIR Process and Decisions

SRI followed the DBIR principles for two years using the following approach:

1. Engage
2. Listen
3. Revise
4. Repeat

DBIR PROCESS FOR POLICY AND CONTEXTUAL ANALYSIS

The two districts that participating in the DBIR process differed on almost everything influencing early design decisions. Because of these challenges, SRI adjusted the original schedule to design the innovation from one to two year, and proposed a pilot study and a field study instead of a two-year field study. Below is a summary of the intended original design, the challenges from the differences between the two school districts, data sources used to understand the challenges, and final solutions. These are separated into tables focused on: standards and curriculum alignment, technology policy, and district and school leadership.

TABLE B1. STANDARDS AND CURRICULUM ALIGNMENT

Original design and intended goal(s)	Challenges	Data sources	Solution
Use Enhanced Units to teach 9th grade biology and 10th grade U.S. History. **Use Enhanced Units to teach 7th grade science and 8th grade U.S. History.**	Curricular topics did not align at the same time during the school year	State standards	Dropped the middle school focus
	Unit lengths varied primarily due to semester length vs. year-long course schedules	Course syllabus for middle school science and U.S. History courses	Forgo grade alignment in high schools
	Philosophically middle school science differed by teaching science courses either year by year or integrated across the types of middle school topics (e.g., earth science, physical science)	Course syllabus for high school biology and U.S. History courses	For high schools, reduced number of units from 8 to 4 units for each subject
	In high school, students can take different paths to satisfy science requirements so courses are taught at different grade levels	Teacher and researcher design group meetings with teachers	Selected the following topical units: Biology: Cells, Ecology, Evolution, Genetics History: Roaring 20's, Great Depression, World War II, Cold War
		Interviews with school principals	

TABLE B2. TECHNOLOGY POLICY

Original design and intended goal(s)	Challenges	Data sources	Solution
Use cell phones to support instruction of the curricular units	Chrome books available on carts but limited use	Teacher and researcher design meetings Principal interviews District interviews, including administrators for information technology and curriculum design Student focus groups	Dropped the idea of using cell phones. Schools were most familiar with using a Google platform, so chose to create an application for use in a Chromebook environment Purchased Chromebooks for the teachers without them
	School web use policies limited how students could support their instruction via the web		
	Teachers used multiple applications within and across schools, with very little overlap		
	Zero tolerance for cell phone use		
	Differences in technology hardware: District purchased tablets for every 9th grade student versus Dell laptop use in other district, although with limited use because of need to check out the carts for use		

TABLE B3. DISTRICT AND SCHOOL LEADERSHIP

Original Design and Intended Goal(s)	Challenges Between Districts	Data Sources	Solution
Principals and district leaders would support the project goals by creating an environment **Teachers were familiar with and using many of the SIM routines planned for use in the study**	Differences in teacher autonomy between the districts: (a) Teachers work independently, setting the sequence, timing and focus of course content. Culture with a lack of accountability for teacher participation (b) Teachers were required to follow the district curriculum and pacing guide, and have students take a state test aligned with the curriculum. District works with teachers to develop course organization and pacing guides No principal engagement vs. Culture of principal involvement to offer encouragement and support accountability On-site SIM consultant vs. Off-site SIM consultant Teachers were less familiar with the SIM routines than originally anticipated	Interviews with district and school administrators Teacher interviews Interviews with district technology and curriculum leaders	Changed the project schedule for the design years from 1 to 2 years. Worked directly with the on-site and off-site SIM coordinators to address teacher use issues. SIM coordinators met with teachers who showed concerns about the use of the Enhanced Unit to understand the challenges and develop a solution. Among the group of teachers, teacher leaders were identified to help reach out to teachers expressing challenges, and develop solutions. District staff and Principal met with the design team in one district to support the project.

SRI stated that the original plans were quite ambitious, but because of their experiences during the DBIR process:

- The final plans are more tempered (and doable)
- Solving the challenges created a collaborative culture among partners to focus on problem solving and compromise
- Resulting in a more realistic fit with many school models rather than just a few, and thus, likely to be more generalizable across schools
- Evaluating the curriculum and technology applications will be the final test – Spring 2018

DBIR PROCESS FOR CREATING THE *EU* AND ASSESSMENTS

Over the two years of designing the curricular units and CORGI, the following activities were completed to engage teachers, administrators, students and researchers,

- **Fifteen full-day or half-day researcher-practitioner design meetings** in which teachers on the design team shared input and feedback on the integrated units, the collaboration strategy and the technology design.
- **Five student focus groups with middle school and high school students** across both districts to collect information on their perception of SIM CERs (routines), and their current use and perception of technology in the classroom.
- **Seventeen student technology pilot focus groups** in which 120 middle and high school students of diverse learning backgrounds piloted early iterations of technology and shared their feedback
- **Fourteen teacher interviews** in which design teachers shared their thoughts on the most effective strategies for training new teachers on the intervention
- **Three interviews with district administrators** to understand district technology policy to develop a technology application to support the curricular units.
- 54 class periods in which U.S. History and Biology teachers piloted the integrated units at multiple stages of development and shared their feedback through a Google survey.

After completing the pilot study, the following changes were made for the field study:

Question 1: Should we limit the number of unit topics (e.g., WWII, Cells)?

Decision 1: The field test will include one practice unit and two study units; coaching will continue during the practice unit but will be a reduced number of hours during the study units. Teachers reported that the level of effort to implement four units was too much to do. This resulted in the decision to use three units. Selection of which unit to drop was determined by analyzing responses to the student end of unit tests, and the team dropped the Cell and Roaring 20s Units.

Question 2: Should we change the *EU* or change the number of devices per *EU*?

Decision 2: The team dropped scientific argumentation.

Question 3: Should we use SCORE?

Decision 3: SCORE was dropped from the field test.

Question 4: Should small schools be included in the field test?

Decision 4: Preference of schools with a minimum of four teacher participants (any combination of U.S. History/biology) to be included in the study for the effort to be cost-effective. Supporting additional SIM coordinators to cover smaller schools would have been cost-prohibitive.

Question 5: Should we include teachers who are new to teaching and/or new to SIM in the field test?

Decision 5: Yes. Not including teachers on improvement plans may be a good idea.

Appendix C. Considerations for Statistical Power

How Large a Sample Do We Need?

We conducted a power analysis to determine how small an impact we could detect given the available sample of classes and students. This is an important part of experimental design, and here we walk through the factors considered.

How Small an Impact Do We Need?

The size of the sample required for a study depends on how small an effect we need to detect. Experiments require a larger sample to detect a smaller impact. It is important to know the smallest potential impact that would be considered educationally useful in the study's particular setting. As a hypothetical example, using percentile ranks as the measure of impact, we may predict that a program of this type can often move an average student 15 percentile points. As a practical matter for educators, however, an improvement as small as 10 percentile points may have value. The researcher may then set the smallest effect of interest to be 10 points or better. Thus, if the program makes less than a 10-point difference, the practical value will be no different from zero. It is necessary to decide in advance on this value as part of the power analysis because it determines the sample size. Conversely, if we had a fixed number of cases to work with, we would want to know how small an effect we could detect—the so-called "minimum detectable effect size" (MDES). Whatever the MDES for a study, it remains possible that effects exist that are smaller than the MDES but that we are unlikely to detect with the sample size available.

How Much Variation is there between Classes?

When we randomize at the class level but the outcome of interest is a test score of students associated with those classes, we pay special attention to the differences among classes in student average scores. The greater the variation in the class averages of student scores, the more classes we need in the experiment to detect the impact of the program. This is because the extra variation among classes adds noise to our measurement which makes the effect of the program, the signal, harder to detect. A summary statistic that is important for the statistical power calculation is the intraclass correlation coefficient (ICC). In technical terms, it is the ratio of the variation in the class averages of students' scores to the total variation in students' scores. A larger ICC means between-class differences in student posttest scores contribute more noise to our program effect estimate. A larger sample of classes is then needed to dampen the noise to acceptable levels. We assume a value of the ICC before the beginning of the study, when conducting the power analysis. (The ICC, like other parameters in the power calculation, reflects our best estimate of what the value is, largely based on compilations of results from other studies. It is not possible to get estimates of these parameters using data from the study at hand until after the study is over.) Certain design strategies are also applied to increase statistical power essentially by accounting for between-class differences that contribute to the ICC. For example, randomizing similar classes within matched pairs removes between-pair differences that contribute noise to the estimate of program impact.

How Much Value Do We Gain From a Pretest and other Covariates?

In order to estimate effects of interest with additional precision, we make use of other variables likely to be associated with performance. These are called covariates because they co-vary with performance on the outcome measure. By including covariates in the analysis, we increase the precision of our effect estimates by accounting for some of the variation in the outcome; that is, by effectively dampening some of the noise so that the signal—the effect of *EU*—

becomes easier to detect. In technical terms, a covariate-adjusted analysis is called an Analysis of Covariance. In our experiments, a student's score on a pretest is almost always the covariate most closely associated with the outcome. Where possible, we adjust for the effect of the pretest. The proportion of variance in the outcome accounted for through modeling covariates is called the "Coefficient of Determination" or R-squared value. In this study we included several covariates, but not the pretest for reasons described in the main body of this report.

How Much Confidence Do We Want to Have in our Results?

We want to be certain that we do not incorrectly conclude (1) that there is no impact when there is one (we want to avoid drawing false negative conclusions), and (2) that there is impact when there is not one (we want to avoid drawing false positive conclusions). Conventionally, researchers have given priority to avoiding false positive conclusions, requiring differences large enough that they would be seen 5% of the time in the absence of an effect before concluding that there is an effect, while at the same time, allowing a conclusion of no effect when in fact there is an effect 20% of the time. For the power analysis, we adhere to these criteria. However, our conclusions reached about the presence of an effect are expressed in terms of levels of confidence rather than as a yes-or-no declaration. As we described earlier in the report, we interpret results in terms of whether they give a lot, some, limited, or no confidence that there is a true impact.

Sample Size Calculation for This Experiment

In this study we had a fixed number of cases to work with. At the outset of the study we projected how small an effect we could detect, the MDES. We assumed 80% power, tolerance for Type-1 error of 5%, 30 classes total, and assuming the pooled analysis for biology and U.S. History combined, an intraclass correlation coefficient (ICC) of 15% (larger than observed from data in the pilot year, but reasonably conservative given that larger ICCs of near .20 are common in educational research) and that use teacher blocks and modeling pretest and other covariates (including students' state test scores from 8th grade and, possibly, scores on the practice unit administered post-randomization) will account for 80% of between-classroom variation, and 40% of within-class variation in outcomes. Under these assumptions the MDES for a 2-level impact analysis is approximately .22 standardized effect size units.

Appendix D. Psychometrics of the Outcome

The tables in this appendix show percent correct, point-biserial correlations, and response rates for individual items on the four unit tests (two unit tests in biology and two unit tests in U.S. History). The figures below each table show the distribution of item difficulty (percent correct) values for each unit test.

TABLE D1. ITEM STATISTICS FOR UNIT 2 BIOLOGY ASSESSMENT

Item	Difficulty (percent correct)	Biserial correlation	Response rate
evoQ1_score	48.10	.37	99.46
evoQ2_score	64.40	.59	100.00
evoQ3_score	61.96	.40	100.00
evoQ4_score	77.72	.44	100.00
evoQ5_score	83.42	.45	99.18
evoQ6_score	77.17	.46	99.18
evoQ7_score	86.68	.31	100.00
evoQ8_score	88.32	.34	100.00
evoQ9_score	86.68	.49	99.46
evoQ10_score	77.99	.43	99.73
evoQ11_score	73.64	.43	99.46
evoQ12_score	78.26	.42	99.73
evoQ13_score	80.16	.27	99.73
evoQ14_score	57.34	.47	99.46
evoQ15_score	70.38	.53	99.73

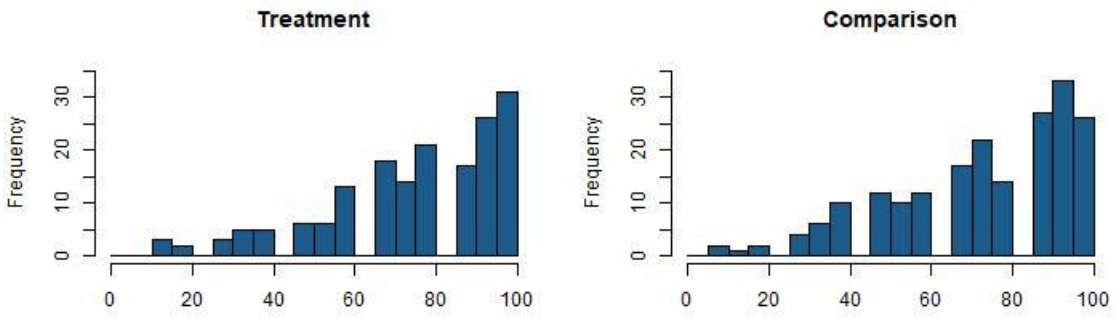


FIGURE D1. DISTRIBUTIONS OF ITEM DIFFICULTY STATISTICS FOR THE UNIT 2 BIOLOGY ASSESSMENT

TABLE D2. ITEM STATISTICS FOR UNIT 3 BIOLOGY ASSESSMENT

Item	Difficulty (Percent Correct)	Biserial Correlation	Response Rate
ecoQ1_score	64.78	.43	100.00
ecoQ2_score	52.42	.41	99.73
ecoQ3_score	51.34	.51	99.73
ecoQ4_score	65.59	.28	99.19
ecoQ5_score	81.18	.38	99.46
ecoQ6_score	83.33	.40	99.46
ecoQ7_score	61.83	.41	98.92
ecoQ8_score	69.35	.35	98.92
ecoQ9_score	56.72	.29	99.46
ecoQ10_score	89.52	.41	99.73
ecoQ11_score	31.72	.28	100.00
ecoQ12_score	68.55	.36	100.00
ecoQ13_score	81.72	.49	99.19
ecoQ14_score	77.69	.45	99.73
ecoQ15_score	79.30	.33	100.00

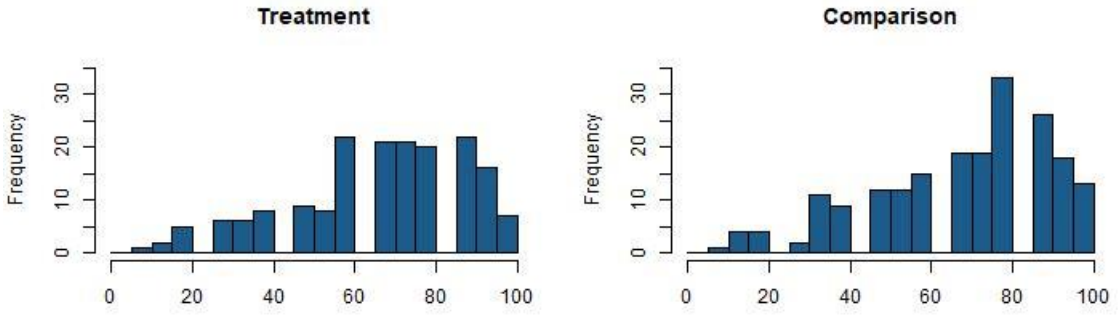


FIGURE D2. DISTRIBUTIONS OF ITEM DIFFICULTY STATISTICS FOR THE UNIT 3 BIOLOGY ASSESSMENT

TABLE D3. ITEM STATISTICS FOR UNIT 2 U.S. HISTORY ASSESSMENT

Item	Difficulty (percent correct)	Biserial correlation	Response rate
ww2Q1_score	46.26	.24	96.92
ww2Q2_score	69.60	.45	98.68
ww2Q3_score	76.65	.49	99.56
ww2Q4_score	88.99	.42	100.00
ww2Q5_score	52.42	.49	98.24
ww2Q6_score	72.25	.43	97.36
ww2Q7_score	64.32	.55	97.36
ww2Q8_score	37.89	.13	98.24
ww2Q9_score	76.21	.45	99.56
ww2Q10_score	30.84	.41	99.12
ww2Q11_score	54.63	.56	99.56
ww2Q12_score	72.69	.46	97.36
ww2Q13_score	70.48	.47	99.56
ww2Q14_score	64.32	.52	98.24
ww2Q15_score	32.60	.47	98.68
ww2Q16_score	32.16	.43	99.56
ww2Q17_score	9.69	.22	97.36
ww2Q18_score	68.28	.48	99.56

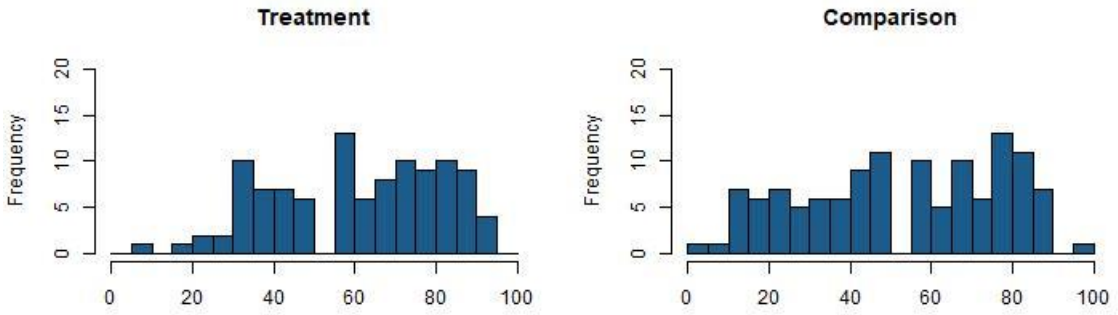


FIGURE D3. DISTRIBUTIONS OF ITEM DIFFICULTY STATISTICS FOR THE UNIT 2 U.S. HISTORY ASSESSMENT

TABLE D4. ITEM STATISTICS FOR UNIT 3 U.S. HISTORY ASSESSMENT

Item	Difficulty (percent correct)	Biserial correlation	Responding rate
cwQ1_score	6.17	-.02	100.00
cwQ2_score	74.01	.25	99.12
cwQ3_score	42.29	.44	97.36
cwQ4_score	32.60	.42	98.24
cwQ5_score	63.44	.50	100.00
cwQ6_score	69.60	.49	100.00
cwQ7_score	80.62	.32	100.00
cwQ8_score	61.23	.57	100.00
cwQ9_score	65.20	.47	100.00
cwQ10_score	10.13	.17	99.56
cwQ11_score	84.58	.41	98.68
cwQ12_score	42.73	.08	99.56
cwQ13_score	8.81	.25	96.92
cwQ14_score	75.33	.33	99.56
cwQ15_score	7.49	.16	100.00

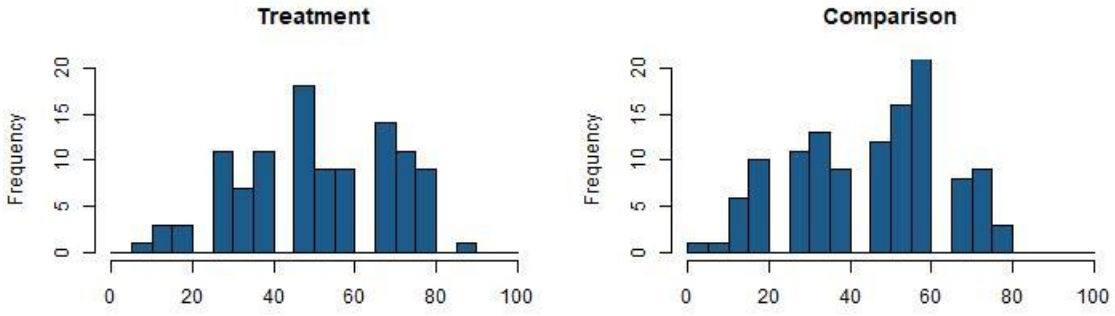


FIGURE D4. DISTRIBUTIONS OF ITEM DIFFICULTY STATISTICS FOR THE UNIT 3 U.S. HISTORY ASSESSMENT

Appendix E. Baseline Equivalence

TABLE E1. TESTS OF BASELINE EQUIVALENCE FOR THE BASELINE AND ANALYTIC SAMPLES

	Baseline			Analytic		
	Combined	Biology	U. S. History	Combined	Biology	U. S. History
Gender Binary Model						
N (students)	642	404	238	619	384	235
Point estimate	0.070	0.057	0.114	0.085	0.063	0.141
Standard Error	.169	0.200	0.319	0.172	0.206	0.315
***p* value**	.680	0.777	0.722	0.621	0.760	0.655
English Speaker Binary Model						
N (students)	642	404	238	619	384	235
Point estimate	0.115	0.700	-0.930	0.140	0.772	-0.950
Standard Error	0.421	0.444	0.843	0.433	0.464	0.848
***p* value**	0.785	0.116	0.271	0.746	0.097	0.264
Is Disabled Binary Model						
N (students)	522	349	173	499	329	170
Point estimate	0.255	0.565	-0.400	0.085	0.420	-0.554
Standard Error	0.845	0.851	2.088	0.913	0.967	2.119
***p* value**	0.763	0.507	0.848	0.926	0.665	0.794
Ethnicity Multinomial Model						
N (students)	642	404	238	619	384	235
Standard Error	0.417	0.521	0.446	0.421	0.529	0.432
***p* value**	0.862	0.872	0.988	0.846	0.865	0.980
Grade Multinomial Model						
N (students)	641	403	238	618	383	235
Standard Error	1.347	1.454	0.685	1.350	1.4453	0.6850
***p* value**	0.589	0.576	0.860	0.609	0.6096	0.8655

Note. We do not report effect sizes for multinomially distributed covariates.

For binary outcomes the point estimates represent the change in log odds of being in the designated subgroup (i.e., being male, being English Language Proficient, being designated as disabled) that is associated with assignment to treatment. For the multinomial model the *p* value tis associated with the test of whether there is difference between conditions in the distribution of individuals across categories, with the cluster correction figured in.

Appendix F. Details of the Approach to Estimating Impacts

Program Impact

The primary question for the experiment was whether, following the intervention, students in *EU* classrooms had higher scores on the end-of-unit tests than students in control classrooms. To answer this question, we analyzed outcomes for the randomized groups. The randomization resulted in two groups that at the outset are statistically equivalent. One receives *EU* and the other one does not. As a result, the average difference between the randomized groups on the posttest is an accurate measure of the program effect plus random error.

We put our data for students, teachers, and classes into a system of statistical equations that allow us to obtain estimates of the effects of interest. The primary relationship of interest is the causal effect of *EU* on achievement as measured by the end-of-unit tests. We use SAS PROC MIXED and PROC GLIMMIX (SAS Institute Inc., 2006) and HLM as the primary software tools for these computations. The output of the analysis process consists of estimates of effects, as well as *p* values that tell us how much confidence we should have that the estimates are different from zero.

We can increase the precision of our effect estimates by accounting for the effects of covariates in the analysis. Therefore, our statistical equations included a series of covariates. We also had to account for the fact that students are clustered by class. We expect outcomes for students who are grouped together to be dependent as a result of shared experiences. We had to add this dependency to our equation in order to prevent artificially high confidence levels about the results. To do this, we modeled a class-level random effect as we describe further in the section below *Fixed and Random Effects*.

Handling Missing Data

To control for potential bias in the effect estimate arising from the covariates having missing values, we used a dummy variable method. With this approach, for each of the covariates that is included in the model, a dummy variable was created. This variable was assigned a value of one if the value of the variable was missing for a given student, and zero, otherwise. The missing values from the original variable were replaced with zero. The dummy method yields effect estimates with less bias than the tolerance threshold set by the What Works Clearinghouse with levels of attrition such as those observed here (this finding is obtained through a simulation study described in Puma et al., 2009). Specifically, the method fares no worse and, in some cases, performs better when compared to other standard approaches, including case deletion and non-stochastic and several stochastic regression imputation methods.

When student achievement outcomes (posttests) were missing, we used listwise deletion and simply dropped the observation from the analysis. This approach to handling missing data is one of several recommended by Puma et al. (2009). In their simulation work, they found that this method produced impact estimates with bias that was smaller than 0.05 standard deviations of the outcome measure (they considered bias in both the estimated impact and its associated standard error).

Potential Mediators

The objective of a mediation analysis is to examine whether an impact of the program on student achievement happens through prior impact on an intermediate outcome such as the use of one or more instructional practices. If an impact is demonstrated on the intermediate variable, and we can also establish an association between the intermediate variable and student achievement independent of the effect of the program, then the intermediate variable may be a mediator of

the impact on achievement. Because we are not randomly assigning cases to levels of the mediator variable, we leave open the possibility that the mediating variables we are examining are proxies for other variables that are the true mediators of the process, but that we have not observed. That is, we cannot be sure of the causal status of the mediator.

We assess mediation whether or not there is an overall impact on student achievement because the mediating path that we are investigating may be one of several, and their effects may cancel when combined, leading to zero overall effect. However, impact on a mediator is necessary (though not sufficient) for that variable to play a mediating role in the impact of *Enhanced Units* on student achievement. As a result of sample size limitations, in this study we simply visually examined differences between Enhanced Units and Control classes in posited mediators of impact on achievement. We did not conduct formal mediation analyses.

Fixed and Random Effects

The covariates in our equations measure either (1) fixed characteristics that take on a finite set of values (e.g., there are only two levels of gender) or (2) a set of characteristics that is assumed to have a distribution over a population and where we treat the values that we measure as though they were a random sample from that larger population. The former is called fixed effects; the latter, random effects. Random effects add uncertainty to our estimates because they account for sampling variation, or the changes we would observe in the outcomes if we re-sampled units from the same population. Fixed effects produce less uncertainty but also limit the extent to which we can generalize our results.

We usually treat the effects of units that were randomized as random effects, so that in the statistical equations, our estimates reflect the degree of uncertainty that comes if we were to draw a different sample of such units from the same population. This allows us to argue for the generalizability of our findings from a sampling perspective. Treating the effects of units that were randomized as fixed forces us to use other arguments if our goal is to generalize.

Using random or fixed effects for participating units serves a second function: it allows us to more accurately represent the dependencies among cases that are clustered together, especially for the clusters randomly assigned to conditions. All the cases that belong to a cluster share an increment in the outcome—either positive or negative—that expresses the dependencies among them. An appropriate measure of uncertainty in our estimate of the program's effectiveness takes into consideration the relative levels of variation *within* and *between* the clusters randomized. All of our statistical equations include an occasion-level error term (Unit 2 or Unit 3), a student-level error term and a randomization-level error term. The variation in these terms reflect the differences we see: (1) between occasions within students, (2) among students within clusters, and (3) across randomized clusters, that are not accounted for by all the other effects in our statistical equation.

The choice of terms for each statistical equation is not rigid but depends on the context and the importance of the factors for the question being addressed. The tables reporting the estimates resulting from the computation will provide an explanation of these choices in table notes where necessary for technical review.

Impact Model Equations

Level 1 (occasion):

$$y_{tjk} = \pi_{0jk} + \pi_{1jk} occasion_{tjk} + e_{tjk}$$

Level 2 (student):

$$\pi_{0jk} = \beta_{00k} + \sum_{r=1}^{m} \beta_{r0k} X_{rjk} + u_{0jk}$$

Level 3 (class):

$$\beta_{00k} = \gamma_{000} + \sum_{s=1}^{n} \gamma_{00s} BLOCK_{00s} + \gamma_{00(n+1)} treatment_{00(n+1)} + \gamma_{00(n+2)} subject_{00(n+2)} + r_{00k}$$

$$\beta_{r0k} = \gamma_{r00}$$

y_{tjk} is outcome t for student i in class k. $occasion_{tjk}$ is a dummy variable indicating if outcome is for Unit 2 ($occasion_{tjk} = 0$) or Unit 3 ($occasion_{tjk} = 1$). X_{rjk} are student-level covariates. $BLOCK_{00s}$ is a dummy variable for block membership, taking value 1 if class k is in block s and 0 otherwise. $treatment_{00(n+1)}$ indicates class randomization status (0 for control, 1 for treatment). $subject_{00(n+2)}$ is used only in the combined analysis of biology and history and is a dummy variable indicating that the result is from a biology class. e_{tjk}, u_{0jk} and r_{00k} are random departures in outcomes from the average at each level of analysis after conditioning on fixed effects in the model.

Appendix G. Reporting the Results

When we run the computations on the data, we produce several results: among them are effect sizes, the estimates for fixed effects, and *p* values.

Effect Sizes

We translate the difference between program and control groups into a standardized effect size by dividing the average group difference by a measure of the variability in the outcome. This measure of variability is also called the standard deviation and can be thought of as the average distance of all the individual scores from the average score (more precisely, it is the square root of the average of squared distances). Dividing the difference by the standard deviation gives us a measure of the impact in units of standard deviation, rather than units of the scale used by the particular test. This standardized effect size allows us to compare the results we find with results from other studies that use different measurement scales. In studies involving student achievement, effect sizes as small as 0.1 (one tenth of a standard deviation) are sometimes found to be important in education. We also report the effect size where we divide the average difference, adjusted for the effects of pretest score and other covariates, by the standard deviation. This is called the 'adjusted effect size'. This adjustment will often provide a more precise estimate of the impact.

Estimates

We provide estimates to approximate the actual effect size. Any experiment is limited to the small sample of students, teachers, and schools that represent a larger population in a real world (or hypothetical) setting. Essentially, we are estimating the population value. When we report an estimate in a table, the value refers to the change in outcome for a one-unit increase in the associated variable. For example, since we code participation in the control group as 0, and participation in the program group as 1, the estimate represents the average difference in outcome that we expect to occur between the program and control group, while holding all other variables constant.

***p* values**

The *p* value is very important, because it indicates how confident we can be that the result we are seeing is not due simply to chance. Specifically, it tells us what the probability is that we would obtain a result with a magnitude as large as—or larger than—the magnitude of the one observed when in fact there is no effect. Roughly speaking, it tells us the risk of concluding that the program has had an effect when in fact it hasn't. This mistake is also known as a false-positive conclusion. Thus, a *p* value of .1 gives us a 10% probability of drawing a false-positive conclusion if in fact there is no impact of the program. This is not to be confused with a common misconception that *p* values tell us the probability of our result being true.

We can also think of the *p* value as the level of confidence we have that the outcome we observe is not simply due to chance. While ultimately depending on the risk tolerance of the user of the research, we suggest the following guidelines for interpreting *p* values.

1. We have a high level of confidence when $p \leq .05$. (This is the level of confidence conventionally referred to as statistical significance.)
2. We have moderate confidence when $.05 < p \leq .15$.
3. We have limited confidence when $.15 < p \leq .20$.

4. We have no confidence when $p > .20$.

In reporting results with p values higher than conventional statistical significance, our goal is to inform the local decision makers with useful information and provide other researchers with data points that can be synthesized into more general evidence.

Appendix H. Fidelity of Implementation, by Subject

TABLE H1. FIDELITY MATRIX FOR THE *EU* KEY COMPONENT 1: BIOLOGY AND U.S. HISTORY TEACHERS RECEIVE SUFFICIENT SUPPORT

Indicator	Operational definition	Source of info/ data collection	Explanation of scoring	Fidelity threshold	Biology Met Fidelity?	History Met Fidelity?
Indicator 1. Teachers receive *EU* PD	PD attendance	Teacher sign-in sheet for 3 days of training; attendance records obtained from coaches	Teacher-level Attended entire 3-day training 0 = did not attend full training, 1 = attended entire training	If a teacher attends the entire 3-day training, he/she will get a score of 1.	7/7 (100%)	5/6 (83%)
Indicator 2. Teachers receive coaching	Frequency and duration that teachers received ongoing coaching	Coach weekly log on *EU* implementation	Teacher-level Total hours receiving coaching 0 = < 8 hours 1 ≥ 8 hours	If a teacher receives ≥ 8 hours of coaching on *EU*, he/she will get a score of 1.	5/7 (71%)	5/6 (83%)
Total teacher-level score for indicator 1 and 2			**1 = teacher attends entire 3-day training AND receives ≥ 8 hours of coaching. 0 = teacher does not receive a score of 1 on both indicators.**	**Total score = 1**	**5/7 (71%)**	**4/6 (66%)**
Indicator 3. Teachers found PD to be useful	Usefulness of PD	Ten-item post-PD survey	District-level (out of 3 districts) Mean score of ratings on the post-PD survey by teachers in the district: 1=strongly disagree, 2=disagree, 3=agree, 4=strongly agree	The mean score on this survey is 3 or above.	n/a	n/a
Criteria for implementing Component 1 with fidelity				At least 85% of teachers have a total score of 1 AND at least 2 of the 3 districts have a mean score of ≥ 3 on the post-PD survey.	Fidelity was not met	Fidelity was not met

TABLE H2. FIDELITY MATRIX FOR THE *EU* KEY COMPONENT 2: BIOLOGY AND U.S. HISTORY TEACHERS USE *EU* UNITS

Indicator	Operational definition	Source of info/ data collection	Explanation of scoring	Fidelity threshold	Biology met fidelity?	History met fidelity?
Indicator 1. Adherence	Reported frequency of *EU* components used	Teacher Implementation Logs, Instructional Practice Survey	Teacher-level Reported use of: (a) Unit Organizer, (b) CORGI at least once with the Unit Organizer, (c) co-construction at least once with the Unit Organizer and once with each of the routines, (d) routines at least once each, (e) teaching background knowledge, (f) using "Cue-Do-Review" (see results below for exact scoring description). For teachers with more than one class, their points were averaged within teacher, across the two classes. In sum, teachers can earn 18 maximum points per unit.	85% (≥ 15.3) of the max possible points If a teacher received an average of at least 15.3 possible points averaged across the 2 study units, the teacher will get a score of 1	6/7 (86%)	3/6 (50%)
Indicator 2. Quality of delivery	Reported combination of *EU* components used	Teacher Implementation Logs	Teacher-level Reported using combination of: (a) Unit Organizer + Corgi + Co-construct and (b) Routine + Corgi + Co-construct (see results below for exact scoring description). For teachers with more than one class, their points were averaged within teacher, across the two classes. In sum, teachers can receive a total of 12 points per unit.	85% (≥ 10.2) of the max possible points ≥ 10.2. If a teacher received an average of at least 10.2 points averaged across the 2 study units, the teacher will get a score of 1	4/7 (57%)	2/6 (33%)
Indicator 3. Usefulness	Usefulness of the *EU*	Instructional Practice Survey	Teacher-level 1 = not at all useful, 2 = less than moderately useful, 3 = moderately useful, 4 = more than moderately useful, 5 = very useful	If a teacher reported an average of 3 or above, he/she will get a score 1.	4/7 (57%)	5/6 (83%)

TABLE H2. FIDELITY MATRIX FOR THE *EU* KEY COMPONENT 2: BIOLOGY AND U.S. HISTORY TEACHERS USE *EU* UNITS

Indicator	Operational definition	Source of info/ data collection	Explanation of scoring	Fidelity threshold	Biology met fidelity?	History met fidelity?
Indicator 4. Student understanding	Students are satisfied that the *EU* helps them to understand the content.	Three-item question on Student Survey	Teacher-level 1=not satisfied, 2=just OK, 3=satisfied, 4=very satisfied; student scores are aggregated to the teacher level	If the average score of students' responses is ≥ 3, the teacher will get a score of 1.	4/7 (57%)	4/6 (67%)
Indicator 5. Student collaboration	Students are satisfied that CORGI helps them to collaborate with their classmates.	Two-item question on Student Survey	Teacher-level 1=not satisfied, 2=just OK, 3=satisfied, 4=very satisfied; student scores are aggregated to the teacher level	If the average score of students' responses is ≥ 3, the teacher will get a score 1.	4/7 (57%)	4/6 (67%)
Criteria for implementing Component 2 with fidelity				At least 85% of teachers have a total score of ≥ 4	3/7 teachers had a score of 4, Fidelity was not met.	2/6 (33%) of eachers had a otal score of , Fidelity was not met

Appendix I. Full Estimates of Benchmark Impact Models

TABLE I1. FIXED AND RANDOM EFFECTS ESTIMATES FOR CONFIRMATORY ANALYSIS OF IMPACT OF *EU* ON BIOLOGY AND U.S. HISTORY COMBINED

Effect	Estimate	Standard Error	t-ratio	d.f.	*p* value
Intercept	36.353	18.008	2.019	14	0.063
Condition	3.090	1.555	1.987	14	0.067
ISMALEN	1.681	1.394	1.206	582	0.228
MSISMALE	11.407	18.641	0.612	582	0.541
ISETHA	1.233	2.481	0.497	582	0.619
ISETHB	-11.722	2.231	-5.255	582	<0.001
ISETHH	-1.639	2.360	-0.695	582	0.487
ISETHI	-4.291	16.988	-0.253	582	0.801
ISETHM	-9.849	5.323	-1.850	582	0.065
ISETHU	-13.940	4.646	-3.000	582	0.003
ISENG	22.182	2.816	7.878	582	<0.001
ISDIS	-14.094	2.491	-5.658	582	<0.001
MSDIS	-5.302	3.281	-1.616	582	0.107
ISGR9	-15.457	17.752	-0.871	582	0.384
ISGR10	-6.533	17.041	-0.383	582	0.702
ISGR11	-6.069	17.346	-0.350	582	0.727
ISGR12	7.454	18.829	0.396	582	0.692
ISUNIT2	7.273	0.673	10.811	536	<0.001
ISBLK18, γ002	0.341	4.227	0.081	14	0.937
ISBLK21, γ003	-10.242	6.433	-1.592	14	0.134
ISBLK22, γ004	-12.415	6.389	-1.943	14	0.072
ISBLK31, γ005	-25.125	6.669	-3.768	14	0.002
ISBLK34, γ006	-11.357	4.637	-2.449	14	0.028
ISBLK35, γ007	3.983	4.990	0.798	14	0.438
ISBLK42, γ008	-8.426	6.614	-1.274	14	0.223
ISBLK43, γ009	-23.926	6.536	-3.661	14	0.003
ISBLK46, γ0010	2.393	4.765	0.502	14	0.623
ISBLK47, γ0011	1.143	4.822	0.237	14	0.816
ISBLK51, γ0012	-7.416	7.111	-1.043	14	0.315
ISBLK52, γ0013	-6.276	7.073	-0.887	14	0.390
ISBLK53, γ0014	12.409	3.992	3.109	14	0.008

TABLE I1. FIXED AND RANDOM EFFECTS ESTIMATES FOR CONFIRMATORY ANALYSIS OF IMPACT OF *EU* ON BIOLOGY AND U.S. HISTORY COMBINED

Effect	Estimate	Standard Error	t-ratio	d.f.	*p* value
ISSUBSCI, γ0015	31.282	7.803	4.009	14	0.001
Random Effect	Variance Component	χ2	d.f.	*p* value	
Residual	130.774				
Student	213.569	2423.518	582	<0.001	
Class	2.744	38.167	14	<0.001	
Standardized Effect Size [a]	.14				
Percentile Standing	6%				

Note. Outcomes were analyzed for 627 student and 30 classes.

[a] The standardized effect size is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE I2. FIXED AND RANDOM EFFECTS ESTIMATES FOR CONFIRMATORY ANALYSIS OF IMPACT OF *EU* ON BIOLOGY

Effect	Estimate	Standard Error	t-ratio	d.f.	*p* value
Intercept	60.629	18.203	3.331	9	0.009
Condition	0.253	1.817	0.139	9	0.892
ISMALEN	-0.854	1.780	-0.480	359	0.632
MSISMALE	15.852	19.334	0.820	359	0.413
ISETHA	-0.034	2.585	-0.013	359	0.989
ISETHB	-11.836	3.015	-3.926	359	<0.001
ISETHH	-7.102	3.028	-2.346	359	0.020
ISETHM	-19.298	7.242	-2.665	359	0.008
ISETHU	-18.155	6.821	-2.661	359	0.008
ISENG	22.403	3.289	6.811	359	<0.001
ISDIS	-12.858	3.001	-4.284	359	<0.001
MSDIS	-5.721	4.447	-1.287	359	0.199
ISGR9	-11.770	17.973	-0.655	359	0.513
ISGR10	-3.236	17.241	-0.188	359	0.851
ISGR11	-5.546	17.746	-0.313	359	0.755
ISGR12	7.101	19.971	0.356	359	0.722
ISUNIT2	6.802	0.878	7.749	330	<0.001
ISBLK18, γ002	7.085	6.926	1.023	9	0.333

TABLE I2. FIXED AND RANDOM EFFECTS ESTIMATES FOR CONFIRMATORY ANALYSIS OF IMPACT OF *EU* ON BIOLOGY

Effect	Estimate	Standard Error	t-ratio	d.f.	*p* value
ISBLK21, γ003	-1.674	5.207	-0.321	9	0.755
ISBLK22, γ004	-4.606	5.192	-0.887	9	0.398
ISBLK31, γ005	-17.487	5.112	-3.421	9	0.008
ISBLK42, γ008	-0.629	5.213	-0.121	9	0.907
ISBLK43, γ009	-16.027	5.260	-3.047	9	0.014
ISBLK51, γ0012	0.502	4.213	0.119	9	0.908
Random Effect	Variance Component	d.f.	χ^2	*p* value	
Residual	137.378				
Student	214.832	359	1460.106	<0.001	
Class	0.063	9	18.036	0.034	
Standardized Effect Size [a]	0.01				
Percentile Standing	0.0%				

Note. Outcomes were analyzed for 391 student and 18 classes.

[a] The standardized effect size is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE I3. FIXED AND RANDOM EFFECTS ESTIMATES FOR CONFIRMATORY ANALYSIS OF IMPACT OF *EU* ON U.S. HISTORY

Effect	Estimate	Standard Error	t-ratio	d.f.	*p* value
Intercept	34.520	21.896	1.577	5	0.176
Condition	6.789	2.396	2.833	5	0.037
ISMALEN	6.084	2.131	2.855	212	0.005
MSISMALE	-1.944	26.230	-0.074	212	0.941
ISETHA					
ISETHB	-9.618	16.053	-0.599	212	0.550
ISETHH	9.829	16.093	0.611	212	0.542
ISETHI					
ISETHM	2.758	17.513	0.158	212	0.875
ISETHU	-4.711	16.664	-0.283	212	0.778
ISETHW	2.710	15.859	0.171	212	0.864
ISENG	27.319	5.339	5.117	212	<0.001
ISDIS	-17.242	4.370	-3.946	212	<0.001

TABLE I3. FIXED AND RANDOM EFFECTS ESTIMATES FOR CONFIRMATORY ANALYSIS OF IMPACT OF *EU* ON U.S. HISTORY

Effect	Estimate	Standard Error	t-ratio	d.f.	*p* value
MSDIS	-8.192	5.003	-1.638	212	0.103
ISGR9					
ISGR10	-21.615	13.329	-1.622	212	0.106
ISGR11	-14.273	11.890	-1.200	212	0.231
ISGR12					
ISUNIT2	8.005	1.040	7.696	205	<0.001
ISBLK18, *γ002*					
ISBLK21, *γ003*					
ISBLK22, *γ004*					
ISBLK31, *γ005*					
ISBLK34, *γ006*	-12.759	5.198	-2.454	5	0.058
ISBLK35, *γ007*	-0.579	6.083	-0.095	5	0.928
ISBLK42, *γ008*					
ISBLK43, *γ009*					
ISBLK46, *γ0010*	-2.579	5.719	-0.451	5	0.671
ISBLK47, *γ0011*	-2.969	5.635	-0.527	5	0.621
ISBLK51, *γ0012*					
ISBLK52, *γ0013*					
ISBLK53, *γ0014*	12.929	3.939	3.283	5	0.022
ISSUBSCI, *γ0015*					
Random Effect	Variance Component	d.f.	χ^2	p-value	
Residual	119.734				
Student	180.743	212	876.665	<0.001	
Class	3.661	5	15.310	0.009	
Standardized Effect Size [a]	0.32				
Percentile Standing	12%				

Note. Outcomes were analyzed for 236 student and 12 classes.

[a] The standardized effect size is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE I4. DESCRIPTION OF VARIABLES USED IN THE BENCHMARK IMPACT ANALYSIS

Condition	Dummy for treatment assignment	=1 if assigned to *EU*, 0 otherwise
ISMALEN	Dummy for gender	=1 if male, =0 if female
MSISMALE	Dummy for missing gender	=1 is gender data missing, =0 otherwise
ISETHA	Dummy for ethnicity Asian	=1 if Asian, =0 otherwise
ISETHB	Dummy for ethnicity Black	=1 if Black, =0 otherwise
ISETHH	Dummy for ethnicity Hispanic	=1 if Hispanic, =0 otherwise
ISETHI	Dummy for ethnicity Native American	=1 if Native American, =0 otherwise
ISETHM	Dummy for ethnicity Mixed	=1 if Mixed, =0 otherwise
ISETHU	Dummy for ethnicity Undeclared	=1 if Undeclared, =0 otherwise
ISETHW	Dummy for ethnicity White	=1 if White, =0 otherwise
ISENG	Dummy for English learner status	=1 if English Language Learner, =0 otherwise
ISDIS	Dummy for Disability status	=1 if Disabled, =0 otherwise
MSDIS	Dummy for missing Disability Status	=1 if Disabled status missing, 0 otherwise
ISGR9	Dummy for member of Grade 9	=1 if home grade is 9th, 0 otherwise
ISGR10	Dummy for member of Grade 10	=1 if home grade is 10th, 0 otherwise
ISGR11	Dummy for member of Grade 11	=1 if home grade is 11th, 0 otherwise
ISGR12	Dummy for member of Grade 12	=1 if home grade is 12th, 0 otherwise
ISUNIT2	Dummy for posttest is for Unit 2	=1 if posttest is for Unit 2, =0 if for Unit 3
ISBLK18, γ002	Block Dummy	
ISBLK21, γ003	Block Dummy	
ISBLK22, γ004	Block Dummy	
ISBLK31, γ005	Block Dummy	
ISBLK34, γ006	Block Dummy	
ISBLK35, γ007	Block Dummy	
ISBLK42, γ008	Block Dummy	
ISBLK43, γ009	Block Dummy	
ISBLK46, γ0010	Block Dummy	
ISBLK47, γ0011	Block Dummy	
ISBLK51, γ0012	Block Dummy	
ISBLK52, γ0013	Block Dummy	
ISBLK53, γ0014	Block Dummy	
ISSUBSCI, γ0015	Dummy for posttest is for Science	=1 if biology posttest, =0 if U.S. History Posttest

Appendix J. Contrast of Additional SIM Instructional Practices

The following figures present treatment-control contrast of additional SIM instructional practices that are not focused on in *EU*.

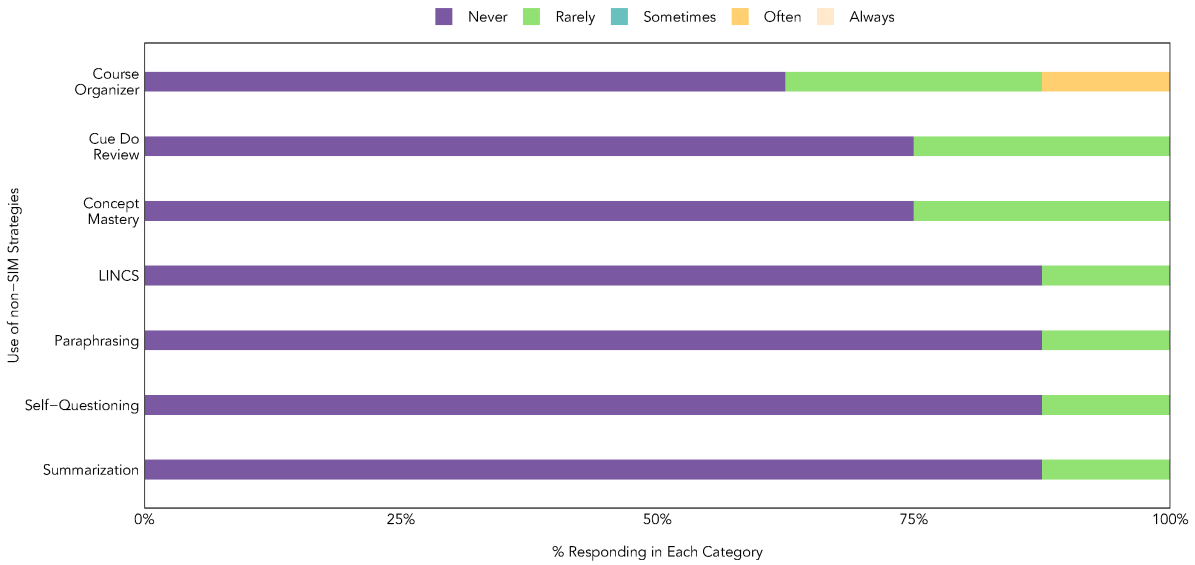


FIGURE J1. FREQUENCY OF USE OF NON-SIM INSTRUCTIONAL PRACTICES, U.S. HISTORY TEACHERS IN CONTROL CLASSES ACROSS UNITS

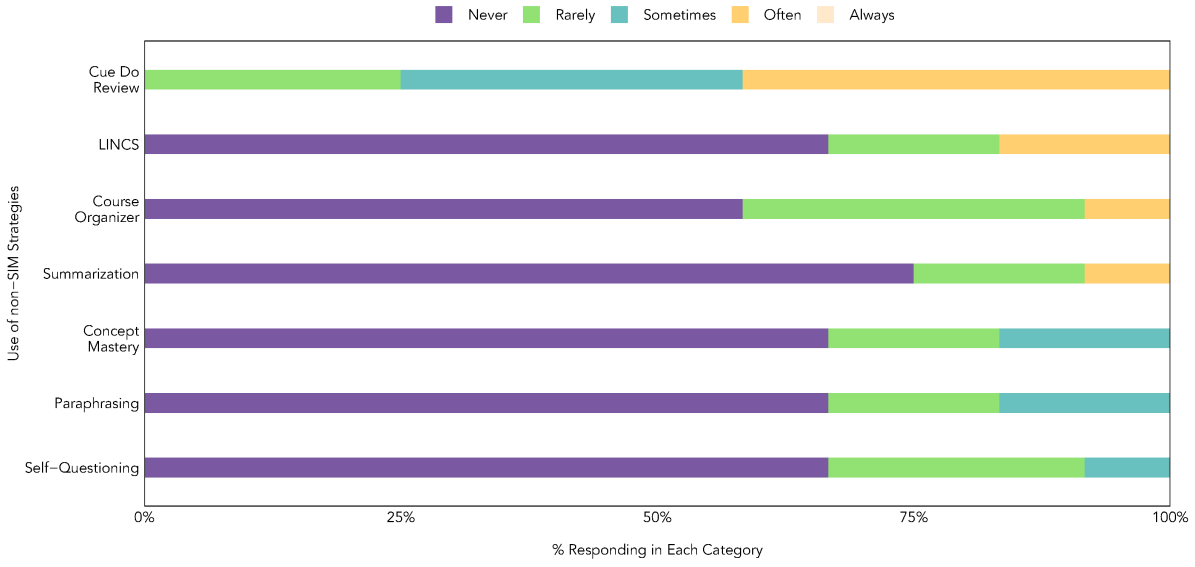


FIGURE J2. FREQUENCY OF USE OF NON-SIM INSTRUCTIONAL PRACTICES, U.S. HISTORY TEACHERS IN TREATMENT CLASSES ACROSS UNITS

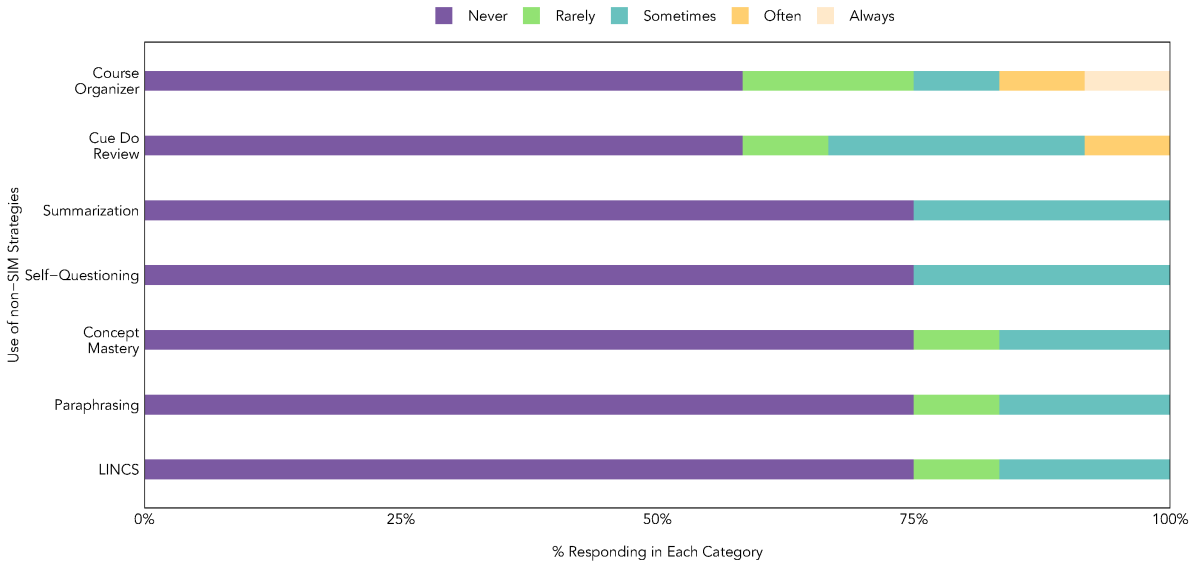


FIGURE J3. FREQUENCY OF USE OF NON-SIM INSTRUCTIONAL PRACTICES, BIOLOGY TEACHERS IN CONTROL CLASSES ACROSS UNITS

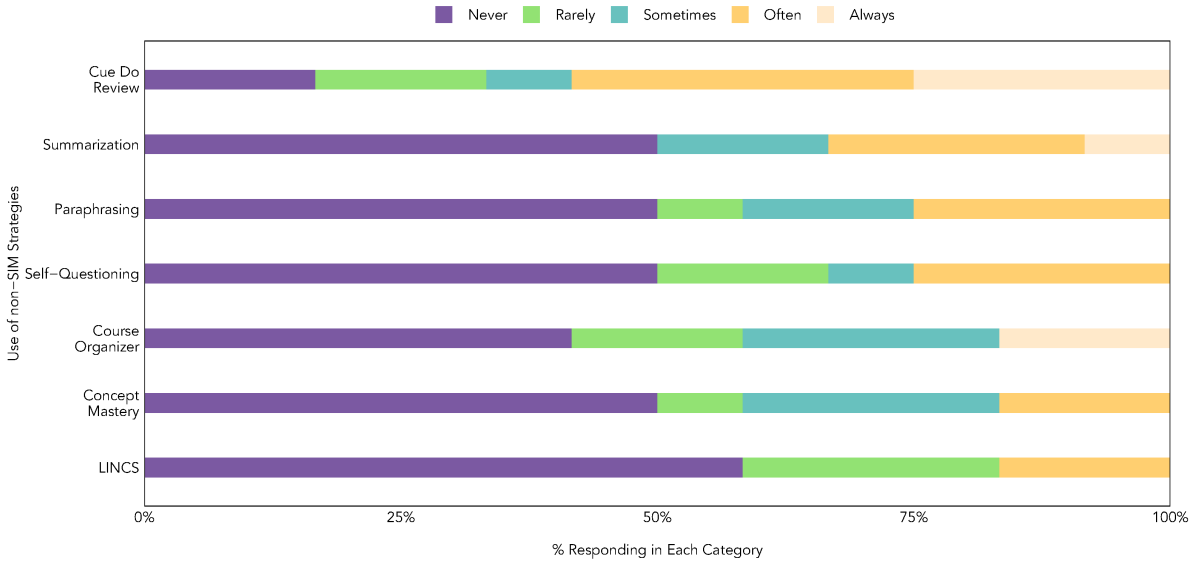


FIGURE J4. FREQUENCY OF USE OF NON-SIM INSTRUCTIONAL PRACTICES, BIOLOGY TEACHERS IN TREATMENT CLASSES ACROSS UNITS

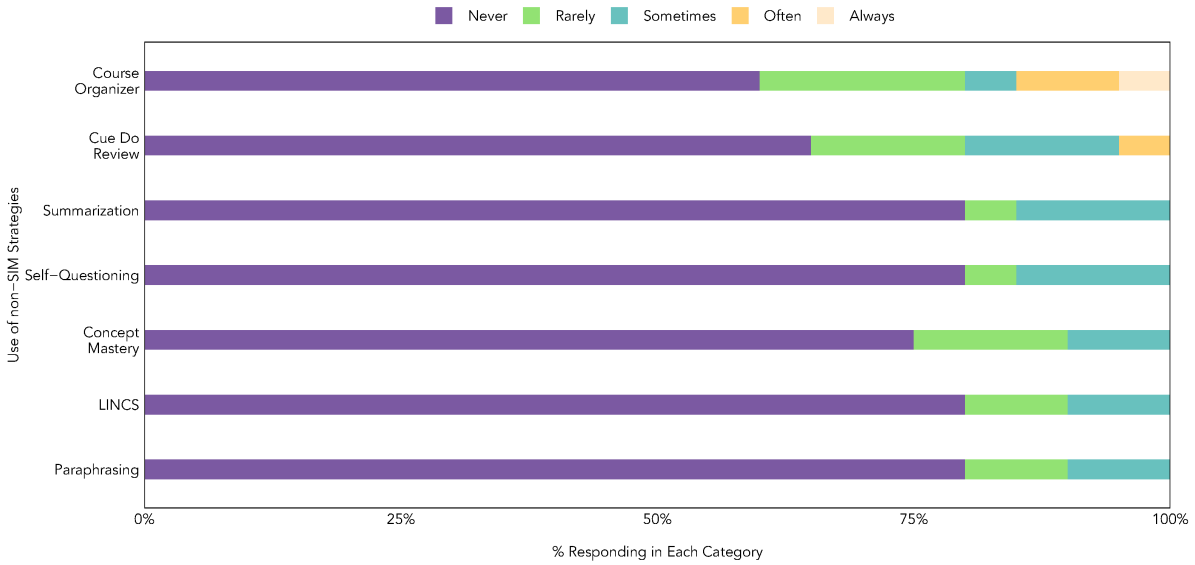


FIGURE J5. FREQUENCY OF USE OF NON-SIM INSTRUCTIONAL PRACTICES, ALL TEACHERS IN CONTROL CLASSES ACROSS UNITS

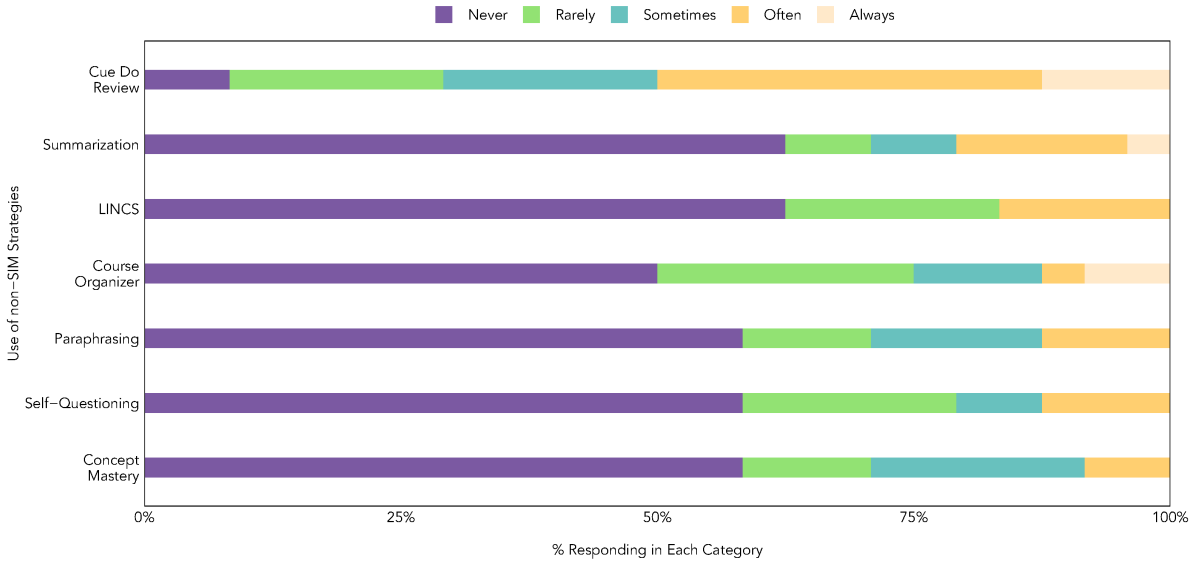


FIGURE J6. FREQUENCY OF USE OF NON-SIM INSTRUCTIONAL PRACTICES, ALL TEACHERS IN TREATMENT CLASSES ACROSS UNITS

Appendix K. The Connection between the Level of FOI and Impact

Table K1 shows the results of progressively adding more covariates into our impact model to account for class-level random sampling variation.

To interpret variation at the classroom level, it is helpful to express it as a proportion of the full sampling variation in the outcome at the individual level; that is, before any adjustment. This ratio of class-level to total variance in the outcome is the intraclass correlation coefficient (ICC). The square root of this quantity allows us to interpret remaining variation at the classroom level in the metric of the standardized effect size; that is, in standard deviation units of the outcome variable.

We observe that the classroom level variance is gradually reduced as we build up the model; from no covariates (Variance=139.79, ICC=.251, sqrt(ICC)=.500) to a fully-parameterized model (Variance=2.63, ICC=.005, sqrt(ICC)=.069). The estimate of this variance component is not statistically significant for all models that include block indicators. After adjusting for effects of covariates, the sqrt(ICC) is roughly .07 standard deviation units. This is small by conventional standards. We can compare this to average impacts of *EU* on the combined, U.S. History and biology outcomes, which were .14, .32, and .01 standardized effect size units.

For exploration, we examined the correlation between block-specific regression-adjusted estimates of the difference between *EU* and control, and each of several fidelity scores. Positive correlations would indicate a relationship between fidelity and impact; while lack of a relationship could indicate either no relationship, or an underpowered test of the relationship. We are interested in whether there are any patterns in the positive direction.

TABLE K1. VARIANCE ESTIMATES FROM SEVERAL MODELS GRADUALLY INCORPORATING FIXED EFFECTS (ANALYSIS OF U.S. HISTORY AND BIOLOG CONSIDERED TOGETHER)

	Variance	*p* value	Variance	*p* value	Variance	*p* value	Variance	*p* value	Variance	*p* value	Variance	*p* value	Variance	*p* value
Class	139.79	<.01	135.6	<.01	68.31	<.01	67.62	<.01	9.98	0.095	9.995	0.094	2.625	0.29
Student	263.92	<.01	263.96	<.01	264.13	<.01	279.5	<.01	280.46	<.01	180.27	<.01	213.6	<.01
Occasion	157.87	<.01	157.86	<.01	157.84	<.01	130.83	<.01	130.84	<.01	130.74	<.01	130.7	<.01
Condition			X		X		X		X		X		X	
Subject (U.S. History vs Biology)					X		X		X		X		X	
Unit							X		X		X		X	
Block									X		X		X	
Unit*Subject											X		X	
Student-level covariates													X	
ICC	0.2508		0.2433		0.1225		0.12131		0.0179		0.0179		0.005	
sqrt(ICC)	0.5008		0.4932		0.3501		0.34829		0.1338		0.1339		0.069	