Designing Activities to Teach Higher-Order Skills: How Feedback and Constraint Affect Learning of Experimental Design

Eli Meir,[†]* Denise Pope,[‡] Joel K. Abraham,[§] Kerry J Kim,[†] Susan Maruca,[†] and Jennifer Palacio[ǁ]

[†]SimBiotic Software, Missoula, MT 59801; [‡]Graduate School, University of Massachusetts, Amherst, MA 01003; [§]Biological Science, California State University, Fullerton, Fullerton, CA 92831; [ǁ]Division of Continuing Education, Harvard University, Cambridge, MA 02138

ABSTRACT

Active learning approaches to biology teaching, including simulation-based activities, are known to enhance student learning, especially of higher-order skills; nonetheless, there are still many open questions about what features of an activity promote optimal learning. Here we designed three versions of a simulation-based tutorial called Understanding Experimental Design that asks students to design experiments and collect data to test their hypotheses. The three versions vary the experimental design task along the axes of feedback and constraint, where constraint measures how much choice students have in performing a task. Using a variety of assessments, we ask whether each of those features affects student learning of experimental design. We find that feedback has a direct positive effect on learning. We further find that small changes in constraint have only subtle and mostly indirect effects on learning. This work suggests that designers of tools for teaching higher-order skills should strive to include feedback to increase impact and may feel freer to vary the degree of constraint within a range to optimize for other features such as the ability to provide immediate feedback and time-on-task.

INTRODUCTION

With the current emphasis on teaching complex, higher-order skills (American Association for the Advancement of Science, 2011; NGSS Lead States, 2013), and a large body of research that students learn such skills better through active-learning approaches (Freeman *et al.*, 2014), it is still an open question what types of active learning are best suited to maximize learning (Behar-Horenstein and Niu, 2011; Freeman *et al.*, 2014). A wide range of classroom activities classified as active learning have been shown effective, but they have many different features (Table 1). The literature contains categorizations of active learning approaches, such as by the degree of scaffolding (e.g., Buck *et al.*, 2008), or along a scale of constructivism (e.g., Arthurs and Kreager, 2017). As designers of educational tools, we consider characteristics that might make an activity effective in classrooms (e.g., McConnell *et al.*, 2017), while not requiring too much instructor effort to be practical for larger classes (Momsen *et al.*, 2010). For instance, computer simulations are known to be effective (Rutten *et al.*, 2012), but it is often unclear exactly which aspect(s) of a simulation-based learning environment makes it effective and studies often lack data on how specific features such as feedback impact effectiveness (Chernikova *et al.*, 2020). For this study, we abstracted three features that have been hypothesized as important.

Some inquiry activities afford students little freedom of choice, which we term here a "constraint" on the students' own exploration (after Scalise and Gifford, 2006). As an example, computer-based questioning systems designed to help students solidify

Anita Schuchardt, *Monitoring Editor*

Submitted Aug 11, 2022; Revised Sep 18, 2023; Accepted Oct 3, 2023

CBE Life Sci Educ March 1, 2024 23:ar1

DOI:10.1187/cbe.22-08-0158

Conflict of Interest Statement: This manuscript describes a study involving a virtual tutorial called Understanding Experimental Design, developed and sold by SimBiotic Software. Eli Meir, Denise Pope, Kerry Kim, and Susan Maruca were all involved in producing the tutorial. Eli Meir, Kerry Kim, and Susan Maruca work for SimBiotic Software, which currently sells the tutorial, and thus have a financial interest in the success of the product. It is not the purpose of the paper to test Understanding Experimental Design, and no promotion of this product to the exclusion of other similar products should be construed.

*Address correspondence to: Eli Meir (eli@simbio.com).

TABLE 1. A selection of activities used in large-enrollment biology classes to introduce more active learning, and some broad characteristics typical of each. See definitions in the text for Constraint and Feedback as used here.

Activity	Constraint	Feedback	Instructor effort (per student)
Polling questions	High	Immediate	Low
Think-pair-share	High	Immediate	Low
Case studies	Intermediate	Immediate to None	High
Written or oral presentations	Low	Delayed	High
Concept-mapping	Intermediate to Low	Delayed to None	High
Intelligent tutoring systems	High	Immediate	Low
Computer-based model construction/simulation exploration	Low	Delayed to None	Low to High
Highly structured simulations	High to Intermediate	Immediate to None	Low
Traditional “hands-on” labs	High to Intermediate	Delayed to None	Low to High

knowledge on a topic (e.g., Urry *et al.*, 2017) are often highly constrained, consisting of multiple-choice or similar format questions, with limited answer options. By contrast, examples of low-constraint activities include building models in a simulation environment (e.g., Klopfer *et al.*, 2009; Bodine *et al.*, 2020), researching and making written or oral presentations, or other activities where there are many paths available for students to take (even if they have highly scaffolded instructions guiding them). The degree and type of constraint, on their own, can affect learning (Meir *et al.*, 2019; Puntambekar *et al.*, 2020).

Another characteristic that differs among active-learning activities is the availability of feedback. Feedback can have a major influence on student learning (Hattie and Timperley, 2007; Shute, 2008), but there are mixed results on when and where feedback is most effective (Kingston and Nash, 2011; McMillan *et al.*, 2013; Van der Kleij *et al.*, 2015; Wisniewski *et al.*, 2020). To help tease apart how feedback influences learning, different authors have proposed categorizing feedback along multiple axes. Proposed categories include immediate versus delayed feedback, the level at which the feedback is aimed (e.g., task vs. process vs. self-regulation), whether the feedback simply provides the correct answer, explains the rationale for that answer, or provides guidance for what the student should try next (Hattie and Timperley, 2007; Shute, 2008; Brooks *et al.*, 2019). Germane to this study, there are indications that the type and timing of feedback can interact with the type of task the student is completing. The optimal timing of feedback (immediate vs. delayed) is still under debate and may be related to whether tasks are aimed at lower- or higher-order thinking (Van der Kleij *et al.*, 2015). Elaborated feedback, where an explanation is provided, has sometimes but not always been shown to be more effective than simply providing the correct answer (Van der Kleij *et al.*, 2015) and its effect may depend on whether task items are highly constrained like multiple choice or lower constraint (LC) constructed response items (Wang *et al.*, 2019). Much of this prior work postulates that features of feedback that make it effective, especially for higher-order tasks, are those that help students reflect on their understanding in ways that help them improve their future performance (e.g., Maier *et al.*, 2016; Brooks *et al.*, 2019). Most importantly to this work, the bulk of previous research has looked at feedback in contexts of either very constrained tasks such as multiple choice questions (Van der Kleij *et al.*, 2015; Zhu *et al.*, 2020), or less commonly lightly-constrained tasks such as constructed responses (e.g., Wang *et al.*, 2019; Zhu *et al.*, 2020), but rarely in the context of tasks with constraint that is intermediate between those such as the type of simulation-based teaching tool we explore here.

Finally, while it does not directly impact student learning, the effort involved in preparing and providing feedback or scoring for each student has a large influence on whether instructors adopt a particular type of activity, particularly for large-enrollment classes, which are typical of many introductory-level science courses (Momsen *et al.*, 2010; McConnell *et al.*, 2017). In Table 1, we summarize these three features for a range of activities that are commonly used in science classes.

In considering how these three characteristics might theoretically influence the effectiveness of activities (Nehm, 2019), we note that the presence, type, and timing of feedback are often dependent on the amount of constraint, as is the per-student instructor effort. That is, providing timely feedback is often only possible when an activity is highly constrained, or at least thoughtfully constrained at some intermediate level (Meir, 2022). Activities where the interaction is highly constrained, such as through multiple choice questions, can easily provide immediate feedback to the learner, with a low level of per-student instructor effort. Low-constraint activities, such as open-ended simulation environments or written or oral presentations typically do not have immediate explicit feedback because there is often no feasible way to provide such feedback, other than in classroom settings where the feedback comes from the teacher responding to student discussion (e.g., Brooks *et al.*, 2019). Instead, they may have implicit feedback (e.g., the student-built model does or doesn’t run or behave as expected), or limited explicit feedback (e.g., the audience asks good questions and/or provides a few thoughts on the presentation), but the bulk of the specific feedback comes to the student with a delay of days or weeks, if ever, once the instructor has a chance to review and assess the work of all of the students (thus, much higher per-student instructor effort).

Many activities fall between these extremes of high constraint with immediate feedback or low constraint with delayed/no feedback. Case studies, for instance, are often guided more lightly than worksheets of practice problems but are more structured than an open-ended research project. Feedback for a case study might be given immediately if on a computer, for instance, if some or all of the questions are in formats that can be algorithmically scored (Clarke-Midura *et al.*, 2018; Magalhaes *et al.*, 2020). Or, feedback might be given with a short delay on a worksheet-based case study when groups of students working in a discussion section might periodically have a conversation with the teaching assistant, with a longer delay if the worksheet is turned in for grading, or even never, if the students complete the worksheet and receive points for it without specific detailed feedback.

There is evidence of both learning enhancement and barriers to learning at different positions on the feedback and constraint axes. Timely feedback often leads to more effective and efficient learning but can also be used by students as a crutch or to game the system by relying too much on feedback rather than thinking through the question themselves (reviewed in: Hattie and Timperley, 2007; Baker, 2011). Different degrees of constraint can similarly be beneficial or detrimental to learning by, for instance, challenging students too little, too much, or just enough given their current skill level (Colburn, 2000; Sweller *et al.*, 2007; Meir *et al.*, 2019; Meir, 2022). Here we ask how these two axes of active learning, feedback, and degree of constraint, may affect learning experimental design, a skill that is complex, difficult, and core to biology (and all sciences).

Experimental design is a difficult higher-order skill

One of the most fundamental skills for students in biology–and indeed all science classes–is designing a good experiment (American Association for the Advancement of Science, 2011; NGSS Lead States, 2013). Experimental process is at the heart of science, yet students often miss important aspects of both the design and implementation of experiments (Dasgupta *et al.*, 2017; Woolley *et al.*, 2018; Pelaez *et al.*, 2022). Because of this, we chose experimental design as our focal skill for this study. Many aspects of experimental design are challenging to students across all levels of study (e.g., Kuhn *et al.*, 2009; Brownell *et al.*, 2014; summarized in: Schwichow *et al.*, 2016; Dasgupta *et al.*, 2017; Pelaez *et al.*, 2017). From this broad literature, we extracted a set of 17 learning outcomes, listed in the Supplemental Materials (Supplemental Table S1), that we used in a backward design process when writing both the learning tutorials and the assessment items in this study. We do not focus further on these learning outcomes here because, while our specific research questions are about experimental design, the purpose of this study is to illuminate how feedback and constraint may affect the learning of higher-order skills more generally.

This study centers on a simulation-based learning tutorial called Understanding Experimental Design (UED) written for students in undergraduate biology classes (Pope *et al.*, 2016). In addition to targeting experimental design learning outcomes (Supplemental Materials), we also designed UED to explicitly test ideas about the role of feedback and constraint in enabling student learning. As authors of open-ended simulation-based learning tutorials that often target higher-order skills, we were frustrated that we were not able to provide immediate, specific feedback to students based on their open-ended explorations using our simulations. This is a common problem as it is hard to provide immediate feedback on LC, open-ended activities, particularly in larger classes. Because of this, much of the research on the effects of feedback is done with more constrained tasks such as multiple-choice questions or memorization of lists (Van der Kleij *et al.*, 2015). We wondered whether a lack of direct feedback reduced student learning efficiency in complex, open-ended tasks. Our premise was that adding some constraints to an open-ended simulation might allow us to provide specific, immediate feedback to students, while still preserving the exploratory aspect of a simulation environment.

Research Questions

This project set out to answer two related questions about how to design effective learning activities for higher-order skills, with our focal skill being experimental design.

1. Does immediate feedback (enabled by constraint) improve student learning of experimental design?
2. Does the degree of constraint (higher or lower) impact student learning of experimental design?

METHODS

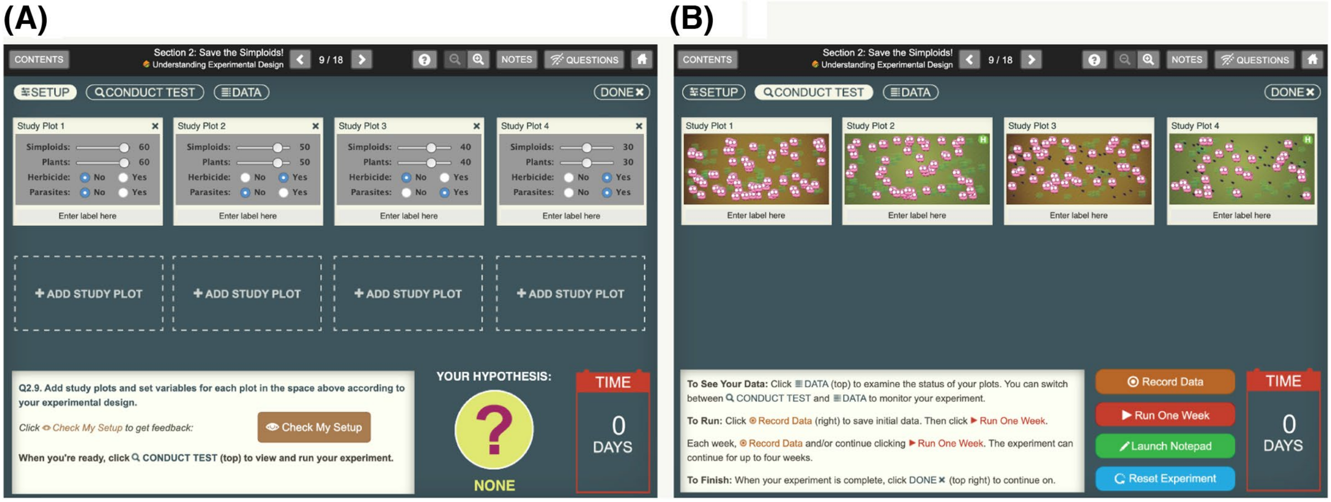
Description of the UED tutorial

The version of UED used in this study was the third major revision of an experimental design tutorial on this study, based on extensive student testing and prior research studies with earlier versions (Abraham *et al.*, 2009). The evolution and justification of choices made in the tutorial and its predecessors are elaborated on elsewhere (Clarke-Midura *et al.*, 2018; Meir, 2022). Here we provide a brief description of UED as it was presented to participants in this study.

Students are given the following scenario. The town of Idyllic has an endemic species called “Simploids” that are beloved by town residents but have recently been getting sick and dying. Students are tasked by the town with doing experiments to discover the cause of the sickness, with two potential causes suspected (parasites and herbicide). The tutorial is split into two sections, which differ in their objectives and level of constraint. This split was based on earlier work on this project (unpublished data) which showed that students’ experimental designs, and their ability to discuss and rationalize those designs with the language of science, were often poorly correlated. We thus aimed one section at teaching the terminology and concepts of good experimental design, and the other on designing, implementing, and interpreting simulated experiments – in other words, the first focused on developing students’ declarative and conceptual knowledge and the second on developing their procedural knowledge (Ambrose *et al.*, 2010, pp. 18–19).

Section 1 is a scaffolded lesson on experimental design that provides students with the building blocks for good experiments and the language used to describe them. Among the concepts covered are systematic variation, scope of inference, independent and dependent (or outcome) variables, treatment and control groups, potentially confounding variables, and replication. Students are assessed on their understanding as they progress through the section with 19 formative assessment items. Most questions in this section use the high-constraint multiple choice format, with five multiple-select, numerical, and other formats that are less constrained than multiple choice but more than open response termed “intermediate constraint” or “IC” (Scalise and Gifford, 2006; Meir *et al.*, 2019).

In Section 2 students design and conduct their own experiments using a simulation of the Simploids. After being guided to choose a hypothesis and plan their experimental design, the heart of the section is an interactive control panel where students design and run their experiment (Figure 1A). They can



	Study Plot 1	Study Plot 2	Study Plot 3	Study Plot 4
Simploids:	60	50	40	30
Plants:	60	50	40	30
Herbicide:	● No ○ Yes	○ No ● Yes	● No ○ Yes	○ No ● Yes
Parasites:	● No ○ Yes	● No ○ Yes	○ No ● Yes	○ No ● Yes
	Enter label here	Enter label here	Enter label here	Enter label here

Q2.9. Add study plots and set variables for each plot in the space above according to your experimental design.

Click Check My Setup to get feedback:

When you're ready, click CONDUCT TEST (top) to view and run your experiment.

To See Your Data: Click DATA (top) to examine the status of your plots. You can switch between CONDUCT TEST and DATA to monitor your experiment.

To Run: Click Record Data (right) to save initial data. Then click ▶ Run One Week.

Each week, Record Data and/or continue clicking ▶ Run One Week. The experiment can continue for up to four weeks.

To Finish: When your experiment is complete, click DONE ✖ (top right) to continue on.

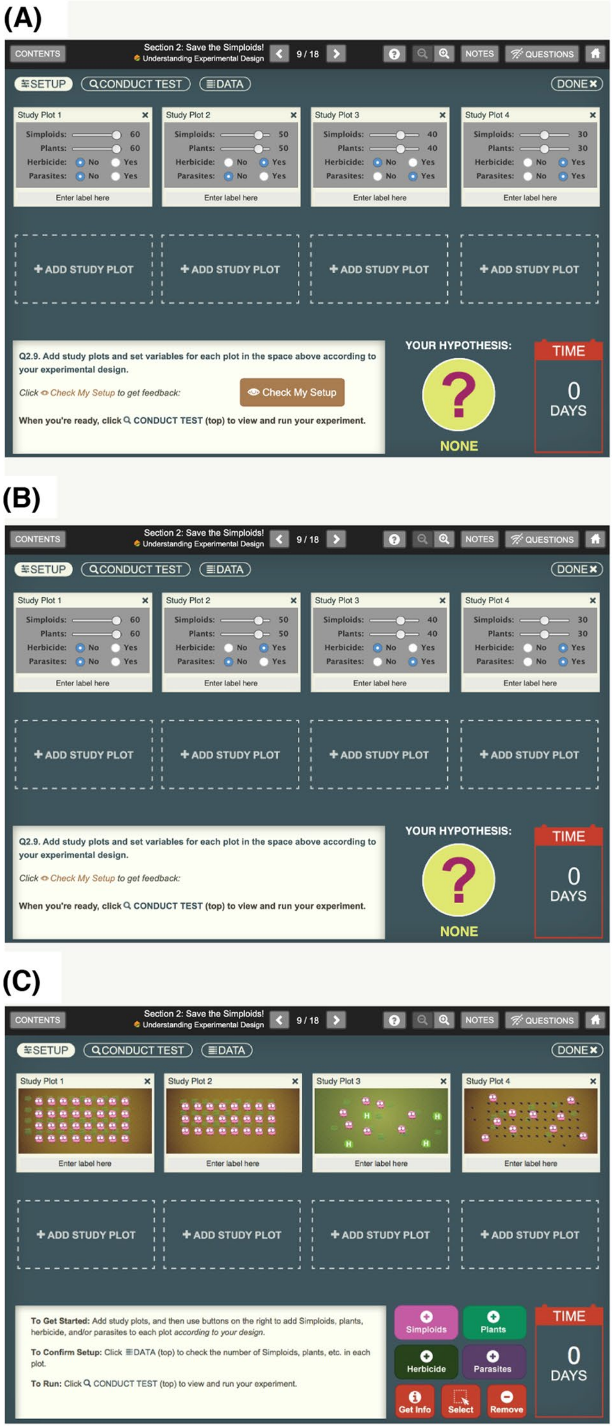
FIGURE 1. Experimental design activity in UED. (A) The panel students use to design their experiment. Students can use up to eight study plots. Each has sliders for selecting number of Simploids and plants, and checkboxes for whether to include herbicide or parasite. A "Check My Setup" button near the bottom provides feedback on the current design. (B) Once ready, clicking "Conduct Test" at the top of the design interface switches to an interface allowing student to run the simulation and collect data. The simulation uses individual-based models. Buttons at the bottom let student choose how long to run and when to collect data.

choose to use up to eight study plots through an "Add Study Plot" button. They must decide how many Simploids and plants (the food of the Simploids) to place in each plot, and whether to include herbicide and/or parasites in each. Once they start running the simulation, they must decide how long to run their experiment (Figure 1B). Students can see the health status of the Simploids throughout the simulation (there are three states – healthy, sick, and dead), and after its completion, an interactive data table lets them view their results. They can adjust the design of their experiment, and complete as many runs as they choose. Each run can last up to 28 d, in 7-d increments. This section also has students answer questions in formats with intermediate degrees of constraint (for instance, constructing sentences from fill-in-the-blank choices [LabLibs: Meir, 2019]) and several short-answer questions on their experimental design plans and conclusions. All questions, except the short-answer format and the interactive data table, provide immediate feedback to students.

Section 2 includes two different experiments, modeling for students that the experimental process is often iterative. After carrying out one or more experimental runs to test one hypothesis for the cause of the disease and drawing a conclusion ("Initial Experiment"), students are asked to consider what they still don't know and design and carry out a second experiment to expand on their knowledge ("Follow-Up Experiment"). For example, they may choose in the Follow-Up Experiment to test the other putative cause of the disease, or test both potential causes simultaneously (they are strongly encouraged in the Initial Experiment to test only one variable). The underlying simulation is complex enough that almost all students have room to learn more about the system after the Initial Experiment.

In the Initial Experiment, students can choose to receive immediate, specific feedback on their experimental design before running their experiment, through a "Check My Setup" button (which they can use as often as they wish). In terms of the major feedback classification systems mentioned in the introduction, this feedback is mostly at the task and process level (how and why to use certain experimental design concepts), elaborated (with explanations), and a mix of what should be done, how to do it, and where to go next. For instance, if a student has no variation between their plots, they receive feedback that reads "To draw conclusions from an experiment you need to create systematic variation so that you can make comparisons between plots. It doesn't look like there is any variation between your plots. In particular, you should vary the independent variable that you specified in your hypothesis." The focus of the feedback is on process rather than on whether the student got the task right or wrong. The words "systematic variation" and "independent variable" link to definitions of those terms (setting expectations), there is an indication of whether the student performed as expected ("it doesn't look like …") and the student is given suggestions for where to go next ("you should vary …"). This feedback has the hallmarks of types of feedback that have been successful in other studies (Brooks *et al.*, 2019).

Our algorithms provide feedback on four aspects of students' experimental design: whether their design systematically varied variables across plots; whether they had appropriate controls for each variable; whether their experiment matched their hypothesis; and whether they had appropriate replication. We did not provide feedback within the experimental design area about two "natural history" related aspects of the experiment: whether they ran the experiment long enough to see the disease progression; and whether they included enough plants to feed all the Simploids in a plot. They should have been able to determine good values for those two parameters from natural history information about the Simploids and the disease progression provided earlier in the section, and from observing the progress of their experiments and examining their experimental results (a form of indirect feedback). The Follow-Up Experiment does not include the "Check My Setup" button, because it is intended



(A)

Study Plot 1 — Simploids: 60; Plants: 60; Herbicide: ● No ○ Yes; Parasites: ● No ○ Yes; Enter label here

Study Plot 2 — Simploids: 50; Plants: 50; Herbicide: ○ No ● Yes; Parasites: ● No ○ Yes; Enter label here

Study Plot 3 — Simploids: 40; Plants: 40; Herbicide: ● No ○ Yes; Parasites: ○ No ● Yes; Enter label here

Study Plot 4 — Simploids: 30; Plants: 30; Herbicide: ○ No ● Yes; Parasites: ○ No ● Yes; Enter label here

Q2.9. Add study plots and set variables for each plot in the space above according to your experimental design.

Click Check My Setup to get feedback:

When you're ready, click CONDUCT TEST (top) to view and run your experiment.

(B)

Study Plot 1 — Simploids: 60; Plants: 60; Herbicide: ● No ○ Yes; Parasites: ● No ○ Yes; Enter label here

Study Plot 2 — Simploids: 50; Plants: 50; Herbicide: ○ No ● Yes; Parasites: ● No ○ Yes; Enter label here

Study Plot 3 — Simploids: 40; Plants: 40; Herbicide: ● No ○ Yes; Parasites: ○ No ● Yes; Enter label here

Study Plot 4 — Simploids: 30; Plants: 30; Herbicide: ○ No ● Yes; Parasites: ○ No ● Yes; Enter label here

Q2.9. Add study plots and set variables for each plot in the space above according to your experimental design.

Click Check My Setup to get feedback:

When you're ready, click CONDUCT TEST (top) to view and run your experiment.

(C)

Study Plot 1 — Enter label here

Study Plot 2 — Enter label here

Study Plot 3 — Enter label here

Study Plot 4 — Enter label here

To Get Started: Add study plots, and then use buttons on the right to add Simploids, plants, herbicide, and/or parasites to each plot *according to your design.*

To Confirm Setup: Click DATA (top) to check the number of Simploids, plants, etc. in each plot.

To Run: Click CONDUCT TEST (top) to view and run your experiment.

FIGURE 2. Three versions of UED. (A) ICWF has sliders and checkboxes for determining contents of each study plot and includes a "Check My Setup" button. (B) ICNF is identical to ICWF but has no "Check My Setup" button. (C) LCNF has students add contents to each study plot with drag and drop for more flexibility and has no "Check My Setup" button.

as a near-transfer assessment where students can apply what they learn from the feedback (both direct and indirect) that they received on the Initial Experiment.

See https://simbio.com/content/understanding-experimental-design for a video introduction to UED. The released version of the UED tutorial (with small modifications from the versions used in this study) is available for evaluation purposes from SimBiotic Software by writing to info@simbio.com and referencing this paper.

Three experimental versions of UED

To separately measure the effects of feedback and of constraint on student learning, our study compared three versions of UED. All versions included the same Section 1 of the tutorial, with different versions of Section 2. One version, which we call IC, With Feedback (ICWF), constrained students in the design activity by only allowing them to select presence or absence of parasites and herbicide in each study plot (using radio buttons), and only allowing addition of Simploids and plants by increments of 10 (using sliders; Figure 2A). This ICWF version includes the "Check My Setup" button which students are free to click at any time while doing the Initial Experiment to receive feedback about their current design. This ICWF version is equivalent to the full version of the tutorial as described in the previous section.

A second version, which we called IC, No Feedback (ICNF), was identical to the ICWF version, except the "Check My Setup" button was hidden, so no feedback was available in the Initial Experiment (Figure 2B). We thus could compare students who completed the ICWF version of the tutorial with students who completed the ICNF version to test for the effect of feedback, without a change in constraint.

A third version which we called LC, No Feedback (LCNF) also was missing the "Check My Setup" button. In addition, rather than radio buttons and sliders, students controlled all four parameters (Simploid population, plant population, parasite, and herbicide) by placing those items in the study plots with the computer's mouse (Figure 2C). They could place multiples of an item with a click and drag (like drawing a rectangle in a drawing program). Students thus had finer control over the number of each item – rather than presence or absence, or multiples of 10, they could place, say, five units of herbicide, three units of parasite, 24 Simploids, and 32 plants, opening up other possible experiments such as testing for dose effects. Students using the LCNF version could create plots with the same parameter settings as were available in the IC versions, as well as many other combinations. Another difference is that the simulation ran in 1-d, rather than 7-d, increments (the maximum duration was still 28 d). Thus, the students in this treatment had more choices for their experimental design, but required a bit more effort per experiment. We note that this is a relatively small reduction in constraint, and many constraints remain (there are still only four variables available, only eight study plots, etc.). By comparing students who complete the ICNF version to those who complete the LCNF version, we are able to isolate the effect of a small change in constraint, without a change in feedback.

An ideal experimental design would also include a LC With Feedback condition. However, we were not able to create that version because our feedback algorithms require a more constrained number of combinations to provide accurate feedback.

The Follow-Up Experiment in each version – the second round of experiments where students are encouraged to test a

TABLE 2. Assessments and populations used to address research questions. See text for descriptions of each assessment and population. Supplemental Tables S1 and S2 are sections 1 and 2 of UED

Comparison	Inference
Assessment: Experimental Design Ability Test (EDCT)	
Samples: Split-Class & Individual	
Pre to Post Paired comparisons, within each population	Effect of UED S1 on declarative & conceptual knowledge of experimental design
Samples: Split-Class & Large-Scale	
Post Mean scores between populations	Generalizability of results
Assessment: In-tutorial experimental designs	
Samples: Split-Class & Individual	
ICWF to ICNF Experimental Design scores	Effect of *feedback* on experimental design skills in UED S2
ICNF to LCNF Experimental Design scores	Effect of *constraint* on experimental design skills in UED S2
ICWF to ICNF to LCNF Biology scores	*Indirect* effects of feedback and constraint on experimental design skills in UED S2
Samples: Split-Class & Large-Scale	
ICWF Experimental Design scores Between populations	Generalizability of results
Assessment: Multiple-Variable Experimental Design Ability Test (MV-EDAT)	
Samples: Individual	
Pre to Post Paired comparisons, all treatments combined	Effect of completing UED on experimental design skills in transfer task
ICWF to ICNF to LCNF Pre to Post change	Effect of *feedback* and *constraint* on experimental design skills in UED S2
Assessment: Interview probing questions following MV-EDAT	
Samples: Individual	
Pre to Post Paired comparisons, all treatments combined	Effect of completing UED on declarative & conceptual knowledge of experimental design

second hypothesis – was similar to the Initial Experiment in most ways, including the level of constraint. However, the "Check My Setup" button was not present in the Follow-Up Experiment in any of the versions. We intended to use students' designs in the Follow-Up Experiments as a performance-based comparison of experimental design ability by treatment.

Student testing of UED and prior versions. We used extensive think-aloud interviews to check the clarity and fidelity to our intent of all UED activities and questions. These were conducted as part of an iterative design-research process, starting with another SimBio module called "Darwinian Snails". That module was first extended with a section on experimental design. Through several more iterations that section was split into its own module, a tutorial on principles of good design was added, and the storyline was changed to discuss the fictional Simploids. We conducted over 80 student interviews throughout this process to gather the data which drove the iterations, and conducted another three specifically with the LCNF version of the tutorial. All interviews were with undergraduate students recruited from introductory biology classes from colleges and universities around Boston, MA, ranging from research universities to undergraduate-focused colleges to community colleges, both public and private.

Measures of experimental design

For the study presented here, we used four assessments of students learning, and three sets of students, to address our research questions (Tables 2 and 3). This section provides a brief overview of the assessments and the next provides an overview of experimental samples.

Screening survey. To stratify students in one of the study sets (see below) by prior understanding of experimental design concepts, we asked each interested participant to fill out an online prescreening survey. In addition to asking several demographic and logistical questions about their availability, the survey contained a nine-question conceptual assessment (see Supplemental Material B) which was used to split students into high, medium, and low performing sets. As this screening survey was used only for this purpose, and for only one set of study subjects, we spent minimal effort on validation and do not discuss those survey results further in this paper.

Experimental Design Concepts Test (EDCT). To assess student understanding of the experimental design concepts and vocabulary that the first section of UED was designed to convey (i.e., declarative and conceptual knowledge), we wrote an assessment we call the Experimental Design Concepts Test (EDCT). While several other assessments on competence in experimental design exist (e.g., Sirum and Humburg 2011; Gobert *et al.*, 2013; Dasgupta *et al.*, 2017; Deane *et al.*, 2017), none published at the time had sufficient coverage of the learning outcomes addressed in UED while also being amenable to autoscoring. The EDCT consisted of 14 multiple choice questions, 11 of which had four answer choices and the last three with two answer choices. All questions were written to address one of the learning outcomes we were targeting (Supplemental Material Section A). The

TABLE 3. Data sources by sample studied[a]

Sample	Individual	Split-Class	Larger-Scale
Versions of UED tested	ICWF ICNF LCNF	ICWF ICNF LCNF	ICWF

MV-EDAT and interview
(Total ED score; ED elements)
	ICWF (*N* = 11);
	ICNF (*N* = 14);
	LCNF (*N* = 14)		

EDCT Pre-Section 1	X (*N* = 41)	X (*N* = 165)	
EDCT Post-Section 1	X (*N* = 41)	X (*N* = 165)	X (*N* = 1292)

UED Section 2 Experimental Designs
(Experimental Score; Biology Score)
	ICWF (*N* = 11);
	ICNF (*N* = 14);
	LCNF (*N* = 14)
		ICWF (*N* = 52);
		ICNF (*N*= 64);
		LCNF (*N* = 44)
			ICWF (*N* = 648)

[a] UED is the UED tutorial, with versions ICWF, ICNF, and LCNF. MV-EDAT is the MV-EDAT and EDCT is the EDCT (see text for details of these tests). Sample sizes in parentheses – sizes listed for each treatment were those used for treatment comparisons.

Supplemental Materials (Supplemental Material Section C2) present several lines of validity evidence that the EDCT was measuring student performance on the focal outcomes.

For the three experimental versions of UED, we placed the EDCT before and after Section 1 to measure learning gains from that highly constrained portion of the tutorial. As Section 1 of UED was identical for all treatments, we also used EDCT results to check for any preexisting differences in performance between treatment groups.

Multiple Variable Experimental Design Ability Test (MV-EDAT). Distinguishing between true learning of experimental design versus learning design within the specific context of the UED tutorial required a performance-based assessment independent of the tutorial. For this purpose, we looked for a pre/post assessment of experimental design procedural knowledge that was open-ended and could capture many of the skills that UED was designed to teach. We started with the EDAT (Sirum and Humburg, 2011) and the Expanded EDAT (Brownell *et al.*, 2014), which prompt students with a real-world scenario and ask them to design an experiment to address the challenge posed in the prompt (e.g., testing the validity of claims that a supplement has a specific impact on human performance). The prompts in those assessments were not well suited to this study, so we built our own derived prompts that we call the MV-EDAT because they include more variables than the original EDAT. There are two versions of the MV-EDAT, one called “Lizard” and the other “Fish”, after the species used in each prompt (see Supplemental Material Section D for more on the prompts and the logic for creating them). Students answered the prompts by drawing and/or writing on the paper that included the prompt.

The MV-EDAT prompts were followed by semistructured interviews to more completely document student declarative and conceptual knowledge. The interview started by asking them to describe the experiment they had designed on paper, and then followed up with questions designed to probe their understanding of experimental design concepts (interview script available on request). To assess their declarative knowledge, some questions asked them to identify elements of their experiment (e.g., “Which is your control group?”); to probe their conceptual knowledge, other questions asked them to explain a concept (e.g., “How do you define control group?”). The interviews allowed us to disentangle procedural knowledge that students draw on when designing an experiment (i.e., the Apply, Analyze, and Create levels of Blooms taxonomy) from declarative and conceptual knowledge that students can cite when prompted (i.e., the Remember and Understand levels of Blooms).

Administration and scoring of the MV-EDAT. Two of the authors [D.P. and J.P.] conducted all the student interviews involving the MV-EDAT. To the extent possible we blinded interviewers and those scoring the interviews to the treatment for each student interviewed. We scored the students’ MV-EDAT experimental designs using both the descriptions they wrote on the paper with the MV-EDAT prompt, and also their verbal descriptions at the start of each interview. For each element of their experimental design where their written and verbal responses differed, they received the higher of the two scores. The experiments described on paper and/or verbally were scored on the presence of six different elements: 1) Uses systematic variation, 2) Addresses hypothesis, 3) Includes replication, 4) Includes variables held constant, 5) Includes dependent variable, and 6) Includes experiment duration. Each of these elements was scored on a scale from 0–2, with 0 being absence (i.e., no systematic variation, no replication, no mention of duration, etc.), 1 being incomplete or partially correct expression of the element and 2 being full and correct expression of the element. We also calculated a total experimental design score by summing all six elements (for a total possible score of 12). Using a randomization test (see below), we found no significant difference between MV-EDAT prompts so we consider the two prompts to be equivalent.

We separately scored eight of the probing interview questions intended to further explore student declarative (four questions) and conceptual (four questions) knowledge of experimental design (we did not analyze all probing questions because the semistructured nature of the interview meant that not all questions were asked consistently of all students). We used a rubric to score responses to the probing questions on the degree of expert-like response, from 0 (no evidence of understanding), 1 (partial evidence of understanding), and 2 (more complete evidence of understanding).

Two team members independently scored every student’s experimental design and response to probing questions using the rubrics described above, and then we discussed and came to consensus on each score. See Supplemental Materials (Supplemental Material D2) for more detail on administering and scoring the MV-EDAT, including measures taken to blind interviewers and scorers to treatments.

Analysis of experimental designs within UED. The final assessment this study relied on was an analysis of the experiments students designed within the UED tutorial. As described in Section 2.1 above, the central activity in Section 2 of UED asks students to design, run, and analyze two experiments, which we designate the Initial Experiment and the Follow-Up Experiment. For both experiments, we logged the design a student made each time they clicked either the "Check Setup" button (to get feedback, available in the ICWF treatment only) or the "Run button" (to run their experiment). We analyzed this data to determine four measures:

1. An "Experimental Score," which combines three factors that is important to a well-designed experiment. Students received one point if their design had systematic variation between experimental plots; one point if they had appropriate control plot(s); and one point if they had full replication of all plots or ½ point for replication of some but not all treatments (e.g., replicating the experimental but not control treatment). The Experimental Score for a student could thus run between 0–3. Crucially, in the ICWF treatment these were all items where the "Check My Setup" button provided specific feedback when their design scored less than one on any of these.
2. A "Biology Score," which combines two factors that are specific to the biological example presented. Students received one point for having sufficient plants in each plot so the Simploids did not starve, and one point for running the simulation long enough that the disease could take its course. The Biology Score for a student could thus run between 0–2. Crucially, the "Check My Setup" button did not offer direct feedback on either of these items, but students had the information about the number of plants necessary, and by observing the course of their simulations, they had information available to infer when they had made these errors (i.e., without sufficient plants, all Simploids in the plot died immediately, and if they had the disease, Simploids first appeared sick before they died).
3. A "Match Hypothesis Score" which tallies whether the experiment the student designed tests the hypothesis they had previously chosen earlier in Section 2. In the Initial experiment, students had a choice of two hypotheses (Herbicide or Parasites as the disease causal factor). In the Follow-Up experiment, they could also choose "a combination of herbicide and parasites." The Match Hypothesis Score was 1 if they varied the variable(s) in their hypothesis and no others, 0.5 if they varied the variable(s) in their hypothesis and others, and 0 if they did not vary the variable(s) in their hypothesis. For the Follow-Up Experiment, we only scored the Match Hypothesis for the Split-Class students, because the hypothesis for the Follow-Up Experiment was an open-response item in the Individual- and Larger-Scale studies.
4. An "Experimental Complexity Score" which measures whether the student attempted to manipulate zero, one, or two independent variables.

See the Supplemental Materials (Supplemental Material E) for more details on scoring of those items.

We looked at the Experimental Score for the *first design* students ran in the Initial and in the Follow-Up Experiments, the *last design* they ran in each, and the *best design* (i.e., highest-scoring) they ran in each. The differences between first, last, and best were small and the patterns were the same regardless of which we chose, so here we report all results using the *first design* students ran in the Follow-Up Experiment, and in the few cases where results differed between experiments, the *best design* students ran in the Initial Experiment. We chose the *best design* in the Initial Experiment as we thought students might learn from earlier runs of the experiment (or the feedback they received, in the case of the ICWF treatment) before running a good design, while we chose the *first design* in the Follow-Up Experiment as we thought by that point in the tutorial, students were less likely to be engaging in exploratory learning and more likely to attempt to directly design the experiment they wanted to run.

Faculty produced good experimental designs. To validate our interpretation of scoring the UED experimental designs, we asked five biology faculty, who we assume have relatively expert knowledge of the experimental design process (especially relative to students) to go through UED, including the experimental design tasks. Four of the faculty-produced experimental designs we would have scored as perfect. One had a design we would have scored as perfect on "Experimental Score" and would have deducted a point on "Biology Score" as they included more Simploids than plants, causing the Simploids to die from starvation.

Given that faculty (our "experts") generally produced designs that our algorithms scored as perfect, we performed statistics on the Experimental Score and the Biology Score by comparing students who received a perfect score versus students who didn't, lumping together all nonperfect scores. This way we were judging the proportion of "expert-like" experimental designs in each sample.

Study samples

This study includes three distinct samples of students. Each sample facilitated the collection of some types of data but not others. To draw robust conclusions required combining the insights gained from each of the three samples. This section describes each sample and the purpose of including them in the study.

Individual comparison of UED versions. To explore the research questions of how immediate feedback and degree of constraint impact student learning, we recruited students from six Boston-area colleges and universities to participate in individual completion of UED along with the MV-EDAT transfer task and one-on-one interviews. We term this the Individual sample. Uniquely among our study samples, this sample provided an opportunity to assess student learning on an experimental design task unrelated to the UED tutorial as we were able to use the time-consuming MV-EDAT and associated interviews. The testing of the Individual sample was structured as follows (also see Table 2, and see above for assessment descriptions).

1. Researchers visited introductory biology courses at Boston-area colleges to announce the study and recruit participants.
   a. Interested students took the Screening Survey online and were invited for in-person interviews. These students were randomized between UED treatments within the low, medium and high strata determined by the Screening Survey.
2. Students came to the study location for an approximately 2-h session, with two parts:
   a. Students completed the MV-EDAT, including the follow-on interview
   b. Students completed Section 1 of UED (the lesson on experimental design vocabulary and concepts), including the EDCT as a pre- and posttest around this first section. Their actions in the tutorial were recorded with screen recording. They were given unlimited time to complete this section of the tutorial. This completed the first interview session.
3. Students returned within a week for a second approximately 2-h session, again with two parts:
   a. Students completed Section 2 of UED (in which they designed and carried out simulated experiments), again with screen recording. They were given unlimited time to complete this section.
   b. After completion, students were given a second version of the MV-EDAT (using a different prompt) including the follow-up interview.

A total of 42 students participated in the interviews, 14 per treatment. Three students in the ICWF treatment were removed from analysis because they never requested any feedback on their experimental designs in Section 2 and therefore could not be used to test for the effects of feedback, leaving a total of 39 students in the study.

Split-class comparison of UED versions. The Individual sample was by necessity limited in size because each interview required extensive time and resources. To expand the data available to address the research questions we conducted a split-class study within an introductory biology class at a western U.S. masters-granting institution. The class consisted of 11 sections of around 24 students ($N = 259$ total), with each section receiving one of the three UED versions. The sections were split across two lecture instructors and five lab teaching assistants. For practical reasons related to the class structure and the software architecture, we needed to assign each section to a treatment rather than randomizing treatments across all students. With only 11 sections, randomizing treatment by section was likely to lead to the introduction of confounding factors. Rather, when assigning the sections to different versions of the UED tutorial, we worked to balance sections across lecture instructors, lab TAs, and lab start times to minimize additional variation between treatment groups (e.g., TAs with two sections would have each section assigned to a different treatment). Data from this sample came from the EDCT and from student designs within the UED tutorial (Table 2).

The UED tutorial was delivered through SimBio's SimUText System, a robust and widely-used software package for distributing simulation-based biology teaching materials (https://simbio.com). The UED tutorial was assigned as homework, with credit given only for completion of the tutorial, not for correctness of responses within the tutorial. Before the UED assignment, another simulation-based module covering topics in evolution (SimBio's Evolutionary Evidence) was assigned, so students had already overcome any logistical challenges to accessing the SimUText System. During the initial subscription process to the SimUText System students were asked for consent to participate in research. When discussing results from the EDCT assessment, we include only data from consenting students who completed all questions of the EDCT both before and after completing Section 1 of UED. When comparing experiments designed in Section 2 of UED between treatments, we include only data from consenting students who completed both the Initial Experiment and the Follow-Up Experiment in Section 2. Students in the ICWF (with feedback) treatment who never requested feedback were removed so data from that treatment only included students who received at least one piece of feedback on their design. Some students completed both EDCT assessments, but not both experimental design activities and vice versa, so the students analyzed for those two data sets overlap but are not identical. We refer to this sample as our "Split-Class" sample (see Table 3 for sample sizes).

Larger-scale testing of UED. To probe the generality of our results, we provided the ICWF version of UED to 27 classes at 14 institutions during the 2016/17 school year (total $N = 1348$ consenting students). This version differed from those used above in that the EDCT was only placed as a posttest after Section 1, but not as a pretest (in the interest of saving class time for instructors volunteering their classes). Of these 27 classes, we dropped three classes that had fewer than 10 consenting students each, along with any student who did not answer all questions on the EDCT, leaving a total of 24 classes and 1292 consenting students. We included all these students when analyzing EDCT scores. Two of these classes did not use Section 2 of UED, and several others had less than 10 students completing Section 2 of UED. Of the remaining 17 classes, only 648 individual students fully completed Section 2. We include only these 648 students when analyzing experimental designs from the Initial and Follow-Up Experiments in Section 2. Data from this sample came from the EDCT posttest and from student designs within the UED tutorial, but unlike the other samples, there was only a single treatment for all students (the ICWF version). Thus, this sample cannot be used to test for treatment effects.

These classes came from a variety of institutions, including one community college, five liberal arts colleges, two masters and six doctoral-granting institutions. Four classes were for upper-level biology students, one was introductory environmental science, and the rest were either majors or nonmajors introductory biology, with the number of consenting students ranging from 11–154 per class. Faculty were recruited through webinars offered via SimBio's mailing list, and only institutions whose own IRB committees approved the study were included in the data used here. Faculty was asked to assign UED for credit in their course, but otherwise was free to use it as they wished. We collected some demographic information from students, including whether they used the tutorial as part of a group or individually, and whether they used it in or outside of class time. We refer to this sample as our "Larger-Scale" sample.

To summarize, we had three experimental samples: Individual, Split-Class, and Larger-Scale, that each provided different subsets of data relating to our research questions. Tables 2 and 3 provide an overview of what data was available from each sample.

Statistics

We collected a large amount of data from each student in the Individual sample, but statistically the sample is small, particularly when divided into the three treatments. We thus chose to analyze data from the Individual sample using a combination of randomization tests and standard parametric tests. Randomization tests can perform better than conventional statistics for small samples and when it is unknown whether the data has a normal distribution (Hesterberg *et al.*, 2003; Craig and Fisher, 2019). We wrote randomization tests in Python (attached: UEDRandomization.py). For comparisons where we ran both parametric and randomization tests, the results were similar and we choose to only report the parametric results for conciseness.

To check for homogeneity of the samples, we compared the ratio of variance on the preassessment scores to that of the postassessment scores and it was suitably low (ratio = 1.66; Craig and Fisher, 2019). We used permutations of the data as is appropriate when looking for significance (Hesterberg *et al.*, 2003) with 10,000 permutations per test. Each permutation took the measured values and redistributed them randomly between students (see below). Where test statistics are two-tailed, we took the absolute value of each calculated statistic. For a statistical significance level of 0.05, we interpret a result as significant if < 5% of the permutation data sets has a test statistic value equal to or higher than the actual data set.

To compare single values between groups, such as comparing MV-EDAT scores between the Lizard and Fish prompts, scores were randomly redistributed between the Lizard and Fish categories during each randomization, and we used the t statistic to compare the groups.

To compare pre- to postassessment scores on the MV-EDAT, we pooled all pre- and postassessment scores, randomly redistributed them between the pre- and postcategories, and then recalculated pre to post differences. Although no statistical difference was found between the Lizard and Fish prompts, we also controlled for differences in prompt difficulty by conducting a second randomization test where values were randomly redistributed only within the same prompt type (i.e., scores on the Lizard prompt were randomly assigned to other Lizard prompts and same for Fish). To control for differences among individual students, we separately tried a third randomization test where pre/post values were only redistributed within each student (i.e., each student had their actual pre- and postscores randomly mixed, but scores were not mixed between students). Neither controlling for interview prompt nor controlling for student differences changed our results, so for simplicity we report only the results for randomizing fully across all students and both prompt types.

We performed all parametric analyses in RStudio version 2023.06.0+421 (R Core Team, 2023). We used repeated-measures ANOVA (RMANOVA) to compare Split-Class and Individual pre- to post performance on the EDCT, and Individual pre- to post performance on the MV-EDAT. We also compared pretest data across the three treatments in the Split Class sample using a single factor ANOVA. We checked all data for normality using Shapiro-Wilk tests and Q–Q plots and, in the case of repeated measures, checked for sphericity using Mauchly's test of sphericity.

Differences on binary variables (student receiving perfect score on Experimental and Biology Scores) were tested using Fisher's exact test in r 3.1.1.

We calculated normalized change scores from pre to post assessments according to Theobald and Freeman (2014). Whenever we ran multiple analyses on the same data we adjusted the alpha level using a Bonferroni correction. We calculated effect size as generalized eta-squared (η^2_g; Lakens, 2013).

RESULTS

Students were comparable in initial experimental design conceptual knowledge across all samples

To check for any differences between treatments within a sample, we compared student performance on the two pre-UED assessments – the EDCT and MV-EDAT. There was no significant difference between EDCT scores of students in the three treatments of either the Individual or the Split-Class sample (F[2162] = 0.766, $p > 0.05$). The sample size precludes us from knowing whether there might have been smaller undetected differences but we have no evidence of any large differences between participants either within or between samples (latter, unpublished data).

We do not have EDCT pretest data for the Larger-Scale sample, but we compared posttest (after UED section 1) EDCT data. Both Split-Class and Individual samples were in the middle of the range of the scores seen with the Larger-Scale classes (Figure 3). We saw no clear trends in the larger-scale EDCT data with class level, institution type, or class size (data not shown).

We also compared preassessment scores on the MV-EDAT between treatments in the Individual sample. Although preassessment scores were a bit higher in the ICWF treatment than the others (4.0 ICWF; 2.6 ICNF; 3.1 LCNF), none of the differences were significant ($p > 0.17$ for all comparisons with randomization test; unpublished data).

Higher constraint first section of UED has almost no effect on students' conceptual knowledge of experimental design

To assess student learning from the higher constraint first section of UED, we compared students' pre- and posttest scores on the EDCT (taken before and after section 1 of UED) using a repeated-measures ANOVA. There were some minor departures from normality in the data, but the assumption of sphericity was met in both the Split Class and Individual Class data. While posttest scores were slightly higher in both the Individual and the Split-Class samples, the differences were very small based on any standard interpretation of effect size (Table 4; Figure 3). Treatment had a small but significant effect across both time points in the Split class, but we did not find evidence of a treatment effect in the Individual Class data. We did not see a significant interaction between the treatment and time in either sample.

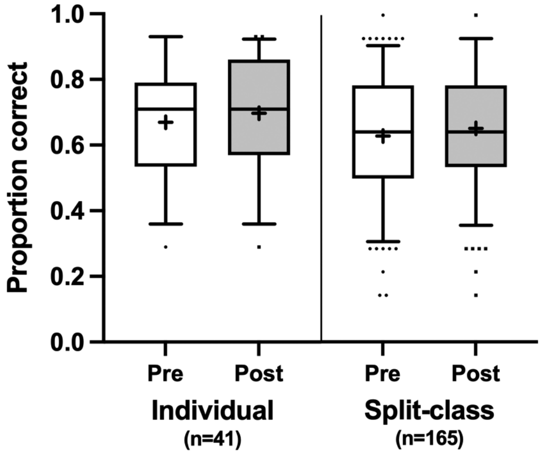


FIGURE 3. Scores on the EDCT from Individual and Split-Class samples. Students in the Individual and Split-Class samples completed EDCT pre- and post-UED Section 1; Larger-Scale sample students completed it post only. Individual and Split-Class samples: Centerlines show the medians; box limits indicate the 25th and 75th percentiles; whiskers extend to the 5th and 95th percentiles; values outside these percentiles are shown as individual points; means are indicated by "+" symbol. The Individual and Split-Class samples were comparable in their pre scores ($p = 0.17$) and both showed very small increases between pre and post.

Student's experimental design skills improve after using UED, as measured by MV-EDAT transfer task

In the Individual sample, we saw overall improvement in students' experimental design skills after completing UED as measured on the MV-EDAT independent experimental design task (Table 4). The data violated both normality and sphericity assumptions. Depending on the sphericity correction, treatment is or is not significant; we took a conservative approach and consider it not significant (Table 4). The experiments that the students designed on paper and described to interviewers were scored on a 0–2 scale on six experimental design elements, for a maximum possible score of 12. The Experimental Design score for students summed across all treatments improved significantly from an average of 4.8 on the preassessment to an average of 7.8 on the post assessment, showing a large effect size of 0.28 η^2_g (or Cohen's $d = 1.0$) (Table 4; Figure 4). Most students (30 of 39) showed an increased score from pre to post. When examined independently, we saw student improvement within each of the three treatments (data not shown) so no one treatment was driving this effect.

Three of the individual experimental design elements in the MV-EDAT – Uses Systematic Variation, Addresses Hypothesis, and Includes Replication – likewise, after Bonferroni correction ($\alpha = 0.008$), showed highly significant improvement between pre- and postassessments for students in all treatments (Figure 5; $p < 0.001$ for each using randomization tests). The other three elements – Includes Variables Held Constant ($p = 0.055$), Includes Dependent Variable ($p = 0.25$), and Includes Experiment Duration ($p = 0.068$) – were not significant at the 0.05 level.

Complexity of designed experiments decreased. To probe how students improved their experimental design skills by completing UED, we looked in more depth at the experimental design elements that showed improvement. The elements Includes Replication and Addresses Hypothesis are both relatively straightforward – that is, after completing UED, more students replicated all treatments in their MV-EDAT designs, and did a better job addressing the stated hypothesis. The change in the Systematic Variation element was more nuanced. We scored students well on Systematic Variation if they changed only one variable per treatment, except for where they explicitly test two variables and their interactions, and they included appropriate control(s). We intentionally included three potential explanatory variables in the MV-EDAT prompts, so that students could choose to test more than one variable in the experiments they designed. They could do this in one of two ways – either design three parallel experiments (one for each variable), or a single experiment that included all three variables. The latter is a more challenging experimental design, because this would require treatments with all combinations of the three variables for a fully balanced design.

On the preassessment, 79% of students designed experiments to test more than one variable (Figure 6). Of these, the majority (77%) attempted to test multiple variables in a single experiment, rather than parallel experiments. This changed in the postassessment, where fewer students (51%) designed experiments to test more than one variable, with only half of

TABLE 4. Summary of repeated-measures ANOVA (RMANOVA) tests for main and interaction effects of treatment and time on student performance on the MV-EDAT and split-class EDCT from individual and split-class populations[a]

Sample	Data	Test		F Statistic	*p* value	Effect size
Individual	EDCT	RMANOVA	Treatment	F(2, 24) = 0.052	>0.05	$\eta^2_g = 0.003$
			Time	F(1, 12) = 9.042	**<0.05**	$\eta^2_g = 0.01$
			Treatment*Time	F(2, 24) = 1.363	>0.05	$\eta^2_g = 0.004$
	MV-EDAT	RMANOVA	Treatment	F(2, 20) = 3.693	**>0.05**[b]	$\eta^2_g = 0.119$
			Time	F(1, 10) = 21.656	**<0.001**	$\eta^2_g = 0.278$
			Treatment*Time	F(2, 20) = 0.347	>0.05	$\eta^2_g = 0.007$
Split Class	EDCT	RMANOVA	Treatment	F(2, 94) = 22.1	**<0.001**	$\eta^2_g = 0.04$
			Time	F(1, 47) = 10.546	**<0.001**	$\eta^2_g = 0.009$
			Treatment*Time	F(2, 94) = 1.678	>0.05	$\eta^2_g = 0.002$

[a] EDCT = Experimental Design Competency Test; MV-EDAT = Multi-variable Experimental Design Ability Test; η^2_g = Generalized eta-squared; bold *p*-values are significant at the 0.05 level.

[b] *p* value corrected due to violation of sphericity; before correction, was < 0.05.

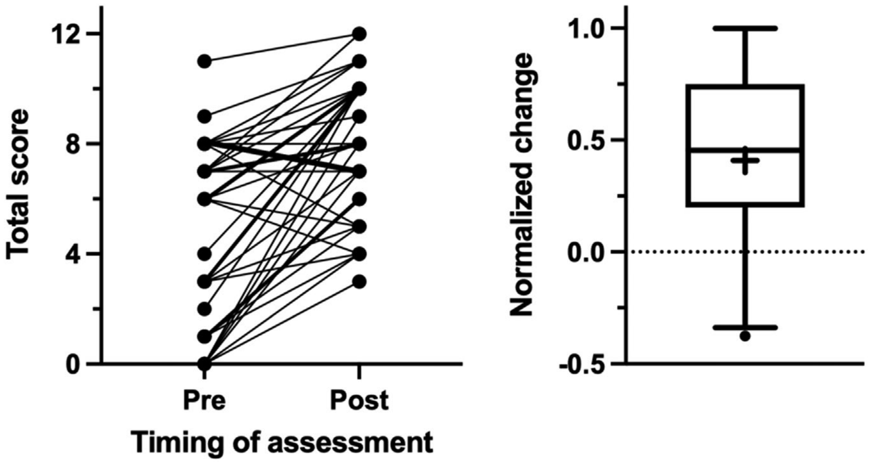


FIGURE 4. Scores on the MV-EDAT in the Individual population ($n = 39$), before and after completing the UED tutorial. Experiments that the students designed based on the MV-EDAT prompts were scored for six design elements on a 0–2 scale, for a maximum possible score of 12. The total score for students in all treatments improved significantly ($p < 0.01$ using randomization test; effect size $\eta_g^2 = 0.278$). First panel: paired pre- and postscores for each student in all three treatments ($n = 39$); 30 students scored higher on the post assessment, two showed no change, and scores decreased for seven students. Line thickness indicates number of students with each pre-post score (e.g., two students' score increased from 1 on the preassessment to six on the post; four students' score decreased from eight to seven). Second panel: normalized change (mean 0.41, indicated by + symbol); centerline at the median; box limits at 25th and 75th percentiles; whiskers extend to 5th and 95th percentiles.

these testing multiple variables in a single experiment. Thus, some of the improvement in the Systematic Variation score is because students chose to test only a single variable in the experiment they designed. But even among those students testing more than one variable in the post assessment, there was evidence of improvement in their experimental design. On the preassessment, the mean Systematic Variation score of those students who attempted to test only one variable was 1.75, compared with a mean of 0.84 for those who attempted to test more than one variable. On the postassessment, the difference was half as much – an average of 1.84 for those who attempted to test one variable, compared with 1.45 for those who attempted to test more than one (Figure 6).

Probing interview questions showed no net change in declarative or conceptual knowledge. After students described the experiment they designed based on the MV-EDAT prompts, we followed up with probing questions that were designed to elicit their declarative and conceptual knowledge. Students did not show any apparent net change in their responses to the eight probing interview questions that we analyzed. For most of the questions, most student answers in the preassessment were scored as one (partial evidence of understanding) or two (more complete evidence of understanding), with very few scores of zero, and the proportions of student scores changed very little in the postassessment. This suggests that the students interviewed came in with a fairly good baseline level of declarative and conceptual knowledge, but this contrasts with their lower level of procedural knowledge, as assessed by the actual experiments they designed in response to the MV-EDAT prompts. For example, most students (over 80% in both pre- and postassessment) could tell us when asked what they would measure, or what variables they would hold constant between groups in their experiment, but about 40% of students did not explicitly include a dependent variable or potentially confounding variables in their description (either verbal or written) of their experimental design, even on the postassessment.

Comparisons between treatments required larger sample and different analyses

The strong and significant overall improvement in performance on the MV-EDAT across all treatments indicates that something about the UED tutorial is working to help students improve on these skills. To address our research questions about the role of feedback and constraint required a comparison between treatments. Ideally, we would have compared changes in MV-EDAT scores between treatments. The small sample size in the Individual sample, though, precludes statistically distinguishing between factors that might be leading to that improvement. So, with this context that a far-transfer assessment shows improved experimental design skill, we focus the rest of our analysis on the experiments students performed within UED. Crucially, the Split-Class sample provides larger data sets from the in-tutorial experimental designs and we can thus use Split-Class data to better test the roles of feedback and constraint.

Feedback affects students experimental design practices

To test the effect of feedback on aiding student learning, we compared the ICWF and ICNF treatments in the Split-Class sample. The only difference between those UED versions was the presence or absence of feedback in the Initial Experiment that students perform. To compare, we examined the number of students who received a perfect Experimental Score on their Follow-Up Experiments, and separately the number who received perfect Biology Scores.

Students in the ICWF treatment were much more likely to have perfect in-tutorial Experimental Scores than those in the ICNF treatment, supporting the inference that feedback aids student learning of experimental design (Table 5; Figure 7B; odds ratio = 4.0).

Students did not receive direct feedback on the components of the Biology Score when making their in-tutorial experimental designs, but they could learn appropriate settings by observing the simulation. In the Split-Class sample, we saw a significant difference on Biology Score between ICWF and ICNF in the Initial Experiment, but not in the Follow-Up Experiment (data not shown).

We see similar patterns in the Individual sample where a higher proportion of students in the ICWF treatment had perfect Experimental Scores than in the ICNF treatment (Table 5;

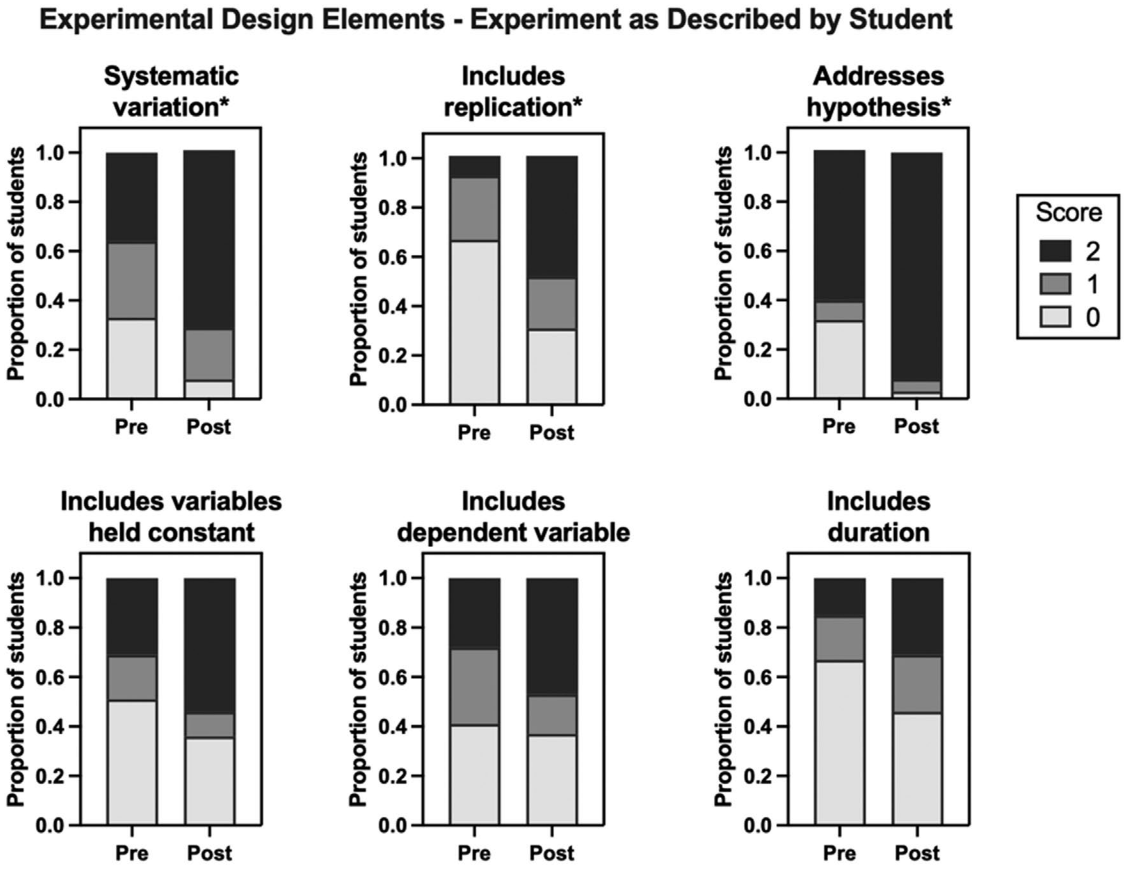


FIGURE 5. Scores for the six individual experimental design elements of the MV-EDAT in the Individual sample ($n = 39$), before and after completing UED. Bars show the proportion of students in all treatments combined scoring 0 (light gray), 1 (medium gray), or 2 (black) on that design element pre and post. The scores for students in all treatments were significantly higher ($p < 0.001$) on the post assessment compared with the pre for three of the individual design elements – Systematic Variation, Includes Replication, and Addresses Hypothesis; the other three elements – Includes Variables Held Constant ($p = 0.055$), Includes Dependent Variable ($p = 0.25$), and Includes Duration ($p = 0.068$) – were not significant.

Figure 7A; odds ratio = 3.4). There were no patterns in that sample for Biology Score.

Change in constraint has little effect on student experimental design practices

To test the effect of constraint on aiding student learning, we compared the ICNF and LCNF treatments in the Split-Class sample. The only difference between these treatments was the degree of constraint (IC; LC) imposed on the experimental design activity. More Split-Class students had perfect Experimental Scores in the ICNF treatment than the LCNF treatments, but the difference was not significant (Table 5; Figure 7B).

When looking at Biology Score in the Split-Class sample, a higher proportion of those in the ICNF treatment incorporated good natural history into their designs than those in the LCNF treatment. This difference was significant on the Initial but not the Follow-Up Experiment (data not shown).

Again, the Individual sample showed a similar pattern. A higher proportion of ICNF students had perfect Experimental Scores compared with LCNF (Table 5; Figure 7A). There were no patterns for Biology Score.

We thus have no evidence that the difference in constraint affects students' learning of core experiment design skills, but there may be an effect on students' ability to properly incorporate details of the experimental system into a good experimental design.

Constraint and feedback have small effects on time-on-task

One might imagine that time on task would be different between the treatments, and this could affect our results. Median time-on-task for the second, open-ended section of UED varied a bit between treatments in the Individual sample (for which we have the most precise data on time spent on the section). Students in the ICWF treatment took a median of 58 min to complete the section, as compared with 61 min for ICNF and 66 min for LCNF. We assume that most of this time difference happened within the experimental design activity, the only part that was different between treatments. This interpretation is supported by the fact that students in the ICWF treatment tended to do fewer experimental runs (combining all runs for both the Initial and Follow-Up Experiments) than students in the other two treatments, in both the Individual (ICWF: 2.73 ± 1.19 SD,

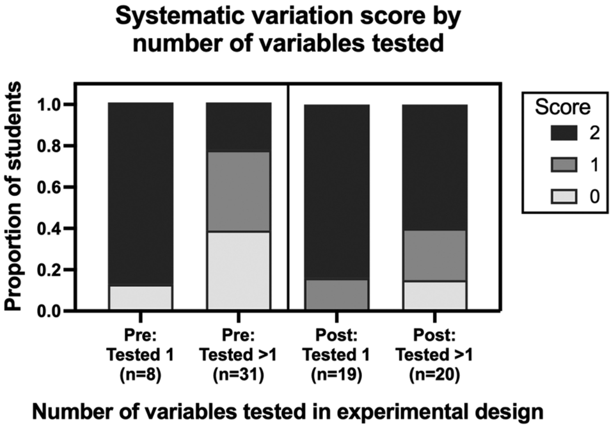


FIGURE 6. Systematic Variation score by number of variables tested in the MV-EDAT. Students are divided into those who designed experiments testing a single variable (Tested 1) or more than one variable (Tested >1), and the bars show proportion of students scoring 0 (light gray), 1 (medium gray), or 2 (black) on Systematic Variation in the pre- and postassessment. More students (n = 31) tried to test multiple variables on the preassessment than on postassessment (n = 20), and those who attempted to test more than one variable scored lower on Systematic Variation. The gap in Systematic Variation scores between students manipulating one versus multiple variables was twice as much in the pre than in the post, suggesting that students learned to test fewer variables and/or when testing multiple variables, to do so more systematically.

ICNF: 3.36 ± 1.45, LCNF: 3.21 ± 1.63) and Split-Class (ICWF: 2.29 ± 0.59, ICNF: 2.52 ± 0.96, LCNF: 2.89 ± 1.48) samples. Thus, higher constraint and feedback likely aided students in completing the key design activity more quickly and enabled them to learn from the feedback directly rather than relying on trial and error.

Split-class and Individual students perform similarly to students in other classes

The Individual sample was relatively small and all came from one metro area. The Split-Class sample was all from one class at one school, albeit in a different metro area. To draw general conclusions from those results, it would be nice to have evidence that those students were not outliers in their experimental design learning. The Larger-Scale sample provides some evidence for this.

The Larger-Scale sample came from 17 classes at institutions around the United States. All students used the ICWF version of UED, so we cannot use these data to probe treatment effects, but we can compare the performance of this sample to the ICWF treatment in the two other samples. In the Larger-Scale sample, 66% of students designed an experiment that received the maximum Experimental Score of three (in individual classes, the proportion of students with scores of three ranged from 30—83%), and 75% received the maximum Biology Score of two (classes ranged from 56–100%) on the first Follow-Up Experiment they ran. The values for the equivalent ICWF treatments in the Split-Class sample (Experimental 69%; Biology 80%) are squarely in the middle of these ranges, and the Individual sample values (Experimental 80%; Biology 80%) are higher but also within the range, indicating that neither of those samples are outliers in their performance on these assessments.

Correlations between assessments are present but vary in strength for different elements

We looked for correlations between the results from the three assessments we used in this study as an indication of whether they were measuring skills similarly. We did this knowing that there are deliberate differences in what each assessment measures.

The MV-EDAT and the in-tutorial experimental designs assess overlapping but not identical skills. Both assess student ability to use systematic variation, to replicate their treatments, and to address their chosen hypothesis, so we compared those specific skills between the two assessments. Of these, the two assessments are most parallel in their presentation and scoring for the skill of replication.

We found significant correlation between MV-EDAT and the in-tutorial experiments for replication. Over half the students (23/39) showed similar skill levels on both assessments (high, medium, or low on both) for replication (p < 0.01, Fisher's exact test).

By contrast, for systematic variation there was little correlation between MV-EDAT and in-tutorial design performance. The MV-EDAT had many more degrees of freedom (e.g., the prompt suggested three possible causal variables, and there

TABLE 5. Comparing In-module experiments between treatments[a]

Comparison	Treatment 1 (perfect/total)	Treatment 2 (perfect/total)	p values	Odds ratio
Split-Class – testing feedback				
Experimental Score Follow-Up Expt.	ICWF (40/52)	ICNF (29/64)	$p < 0.01$	4.0
Biology Score Initial Expt.	ICWF (46/52)	ICNF (43/64)	$p = 0.008$	3.7
Biology Score Follow-Up Expt.	ICWF (45/52)	ICNF (49/64)	$p = 0.23$	2.0
Split-Class – testing constraint				
Experimental Score Follow-Up Expt.	ICNF (29/64)	LCNF (16/44)	$p = 0.43$	1.5
Biology Score Initial Expt.	ICNF (43/64)	LCNF (19/44)	$p = 0.02$	2.7
Biology Score Follow-Up Expt.	ICNF (49/64))	LCNF (27/44)	$p = 0.13$	2.0

[a] Comparison is of students who received perfect scores on each component of experimental design, shown as a ratio to total number of students per sample. Feedback tests compare ICWF to ICNF treatments; Constraint tests compare ICNF to LCNF. Comparisons use Fisher's exact test.

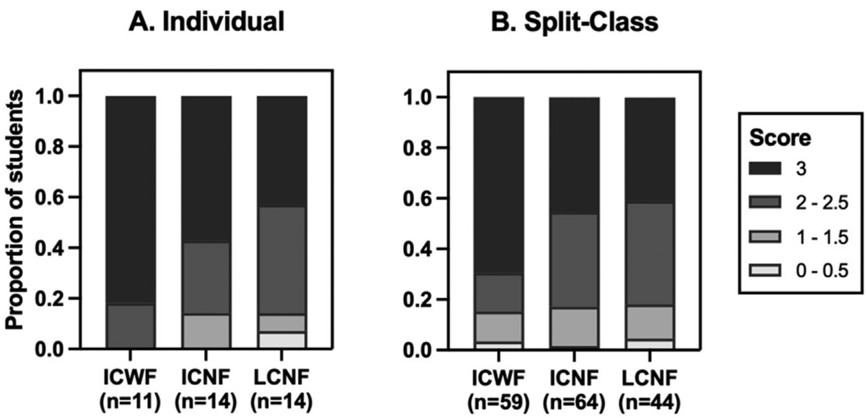


FIGURE 7. Experimental Scores for the Follow-Up Experiments designed by students in Section 2 of the UED tutorial, by treatment. The components of the Experimental Score were: systematic variation (score 0 or 1); appropriate controls (score 0 or 1); and replication (score 0, 0.5, or 1), for a total possible score of 3. Effect of feedback: a higher proportion of students in the ICWF treatment achieved perfect Experimental Scores of 3 compared with the ICNF treatment in both the Individual (A) and Split-Class (B) sample. The difference is significant in the Split-Class sample ($p = 0.0006$ with Fisher's exact test; odds ratio = 4.0) and follows the same pattern in the Individual sample (sample was underpowered for statistics; odds ratio = 3.4). Effect of levels of constraint: more students in the ICNF treatment achieved perfect scores compared with the LCNF treatment, but the difference was not significant.

were no suggestions about how to group or house the animals being tested).

There was also no significant correlation for whether students could fully match their hypothesis to their experimental design ($p = 0.22$, Fisher's exact test, comparing pretutorial MV-EDAT and Match Hypothesis score for the Initial Experiment). This is somewhat confounded, though, because in the tutorial, the page where students choose their first hypothesis is separated by several pages and multiple activities from where they conduct their Initial Experiment, while in the MV-EDAT there was no separation. We do note that in both the MV-EDAT (pretutorial) and the Initial in-tutorial experiment, most but not all students were able to fully match their hypothesis to their experiment in both Individual and Split-Class samples (MV-EDAT pretutorial = 61%; Individual Initial Experiment = 50%; Split-Class Initial Experiment = 58%), indicating some consistency in this skill between the assessments.

The questions in the EDCT were written to target the learning outcomes focused on in Section 1 of the UED tutorial, which were focused on declarative/conceptual knowledge of experimental design, mostly at the lower (Remember and Understand) levels of Blooms Taxonomy (Remember and Understand), while the experiments in Section 2 were designed to assess students' procedural knowledge of these concepts (i.e., the Apply and Create levels of Blooms). We compared total EDCT posttest score against a sum of each student's Experimental, Biology, and Match Hypothesis scores on their experiments in Section 2. As scores on the in-tutorial experiments varied by treatment, we did separate comparisons against each treatment, using data from the Split-Class samples. Correlation between the EDCT and in-tutorial scores ranged from low ($r^2 = 0.3$) for ICWF to virtually nonexistent for the other two treatments.

In the Individual sample, we compared posttest EDCT and MV-EDAT. There was no correlation between scores on those assessments.

DISCUSSION

Active learning is widely accepted as good practice in science education after so many studies have shown active approaches to be superior to passive approaches in teaching (Freeman *et al.*, 2014). But simply claiming to use active learning practices is not guaranteed to result in improved learning (Andrews *et al.*, 2011), and certain active learning practices are more effective than, or better when combined with, others (e.g., Nehm *et al.*, 2022). The questions now have moved beyond comparing active and passive learning to research on what particular aspects of an active learning approach lead to effective learning. Pushing this research forward is particularly important for complex concepts and skills which are harder to measure and thus less likely to have developed clear guidance. In this study, we aimed to probe what design features of an activity to teach a complex biological skill – experimental design – led to increased learning. In particular, we look at the effects of feedback and constraint on learning in a digital inquiry-driven tutorial called UED.

The UED tutorial is effective at teaching experimental design skills, with large learning gains among college biology students (η^2_g effect size = 0.28; Cohen's *d* effect size = 1.0) as measured by our independent assessment of experimental design ability, the MV-EDAT (derived from the EDAT; Sirum and Humburg, 2011). We do not have a comparison group of students who did not use UED in this study, so we draw no conclusions about whether UED is more or less effective than an equivalent use of time with another experimental design activity. We do note, though, that the learning gains we measured compare favorably to other activities designed to teach the experimental design process. Using the similar Expanded-EDAT assessment, Brownell *et al.* (2014) report an effect size of Cohen's $d = 0.36$ from a one-class period pencil-and-paper experimental design activity, while semester long courses with a focus on experimental process report effect sizes (measured with Cohen's *d*) from 0.38 (Abdullah *et al.*, 2015) to 0.99 (Shanks *et al.*, 2017). Thus, it seems likely that UED is at the least equivalently effective to other possible activities designed to teach similar skills.

Within the context of demonstrating that UED is an effective activity for learning experimental design skills, our study tried to tease out what features were most responsible for these learning gains, focusing on constraint and feedback. The metrics we have for this are not perfect. Ideally, we would have used our MV-EDAT data to compare treatments, but the interviews were too time-intensive to provide large enough sample sizes to allow for robust comparisons. We, therefore, draw many of our conclusions from comparing the experiments that students designed within Section 2 of the tutorial in the three treatments of the Split-Class sample. We argue that between-treatment differences in student performance on the in-tutorial experimental design tasks in this sample likely reflect differences in learning. This argument is supported by a correlation between the in-tutorial experimental designs and the MV-EDAT data on at least one measure (Replication), and some similarity in a second measure (Match Hypothesis), indicating that these two assessments measure related, though not identical, skills.

Feedback contributes to learning within an intermediate constraint activity

Our Split-Class data clearly shows students who received specific feedback on the experimental design task designed better experiments in UED, as measured by their Experimental Scores, which summed their score for systematic variation, appropriate controls, and replication of all treatments in experiments designed within the tutorial. The Experimental Score was higher in the ICWF treatment, where students received feedback, than in the ICNF treatment, where they did not (Figure 7). We saw this higher Experimental Score in the Follow-Up Experiment where neither treatment provided any feedback, so it was not just a function of students responding directly to the immediate feedback they received, but represented learning that lasted at least a short time and transferred to a similar activity.

Students did not receive feedback on the aspects of their design that went into the Biology Score (supplying enough plants to feed the Simploids and running the experiment long enough to see the disease progress), and although the students in the ICWF treatment had higher Biology Scores in the Initial Experiment, in the Follow-Up Experiments there were no differences between the ICWF and ICNF treatments. This lack of difference on the Follow-Up Experiment supports the impact of feedback on performance, because students in the two treatments ended with equivalent scores for an aspect of the experimental design where we did not provide direct feedback.

That feedback helps students is not a surprise, especially on learning higher-order tasks (Van der Kleij *et al.*, 2015). While our study was not designed to test different types of feedback, our effect size was large compared with many other studies of feedback effectiveness (Van der Kleij *et al.*, 2015; Wisniewski *et al.*, 2020). These results thus lend support to previous research indicating immediate, elaborated feedback that includes information on what to do next is effective (e.g., Brooks *et al.*, 2019; Wisniewski *et al.*, 2020), especially for higher-order tasks (Van der Kleij *et al.*, 2015). More novel to this study is showing those characteristics of feedback remain effective for automated feedback on higher-order, intermediate constraint activities such as the simulation-based activities in UED, where providing feedback is challenging and few previous studies have been conducted. The idea for the UED experimental design activity originated with an earlier activity on natural selection called Darwinian Snails (Abraham *et al.*, 2009, Clarke-Midura *et al.*, 2018) which had a much less constrained experimental design activity without any feedback. In our iterative testing of various learning modules with individual students and classes, we have found that by adding some constraints to an activity while still retaining much of the open-ended nature of a learning environment, as we have done in UED, we are able to devise algorithms to automatically provide specific feedback (Meir, 2022). Our results here show this is worth doing as the feedback clearly improves student performance and, we infer, student learning. In particular, there is a larger difference in student scores on the transfer task (the MV-EDAT) on skills for which explicit feedback was provided in the tutorial (e.g., Systematic Variation, Includes Replication, and Addressing Hypothesis) compared with those for which they did not receive feedback (Figure 5).

Small changes in constraint have little effect on learning

When teaching complex scientific skills, how much freedom should one provide students within a learning environment? There are arguments in favor of both heavily constraining the student experience to make it easy for students to perform the skill ("structured inquiry" as defined by Colburn, 2000; Sweller *et al.*, 2007), and on the other extreme providing a largely discovery-based approach where students are given a space in which to explore and discover largely for themselves ("open inquiry"). Many activities provide constraint intermediate between those extremes, but even within the intermediate range, it's not clear where on that axis learning best takes place.

Here we tested how varying constraints within an intermediate range might affect learning, and for the most part, we find little effect. We would consider both ICNF and LCNF treatments to be intermediate in degree of constraint compared with other activities we and other groups have designed. Within this range, we see only minor effects on learning. While there was a trend towards Experimental and Biology Scores being lower in LCNF, aside from Biology Score in the Initial Experiment, the difference was not significant. It is possible that with a larger sample those trends would have risen to significance, but from our results we conclude that even if an effect is there, it is not large.

This contrasts with other data comparing a much broader range of constraint on question types. In a separate, related study, we compared questions written in open-ended (short answer) versus intermediate constraint formats, for instance filling in blanks of sentences from predefined sets of words and phrases ("LabLibs"; Meir *et al.*, 2019). There, we found that student learning increased when they were asked questions in an intermediate constraint format compared with the same material using open-ended questions, in some cases even without specific feedback on their answers. Other research groups have also opted for intermediate levels of constraint in both learning and assessment environments (e.g., Blanchard *et al.*, 2010; Gobert *et al.*, 2012). Anecdotally, when we watched students use the much lower constraint experimental design activity in the Darwinian Snails module from which UED evolved, it appeared that many (perhaps most) did not take full advantage of that environment to truly explore. This aligns with data showing students using intermediate constraint questions can express their thinking with more clarity than in essay questions (Meir *et al.*, 2019). Constraint may guide students towards more productive thinking and exploration in open environments. Our results here are consistent with previous work showing intermediate constraint activities are helpful in promoting learning, but do not offer much evidence in favor of the hypothesis that level of constraint matters.

Instead, we take these results as an invitation to consider other factors when developing learning tools. Rather than worrying about direct effects on learning of different levels of constraint, the primary considerations within the broad intermediate region may be indirect effects on other aspects of the environment, such as ability to provide feedback and learning efficiency. To promote discovery learning, one might try to target the least constrained environment for which one can still devise algorithms to provide good feedback. In this light, as algorithms for providing feedback improve, it would be interesting to repeat the experiments we have done here comparing intermediate and low constraint environments where both have immediate, specific feedback. In the other direction, we note that students in the LCNF treatment took longer to complete the tutorial, without any evidence of greater learning. Assuming time on task tracks with extraneous cognitive load, one might also reasonably increase the constraint on the environment to maximize learning efficiency (Paas and Van Merrienboer, 1993). Increasing constraint enough to provide automated feedback and scoring benefits instructors as well, allowing them to use activities in larger classes with less effort (Table 1). Thus instructional designers might manipulate constraint higher or lower to maximize learning based on other factors dependent on the constraint, without worrying about degree of constraint itself impacting learning.

Results may apply across a broad range of undergraduate students

While we were not able to conduct controlled comparisons in more than one class, we were able to compare data from the ICWF version of UED gathered from a broad range of classes across the United States. There were wide variations among student scores between classes, as one might expect. But the results from our Individual and Split-Class samples fit well within the range seen in other classes on both in-tutorial experimental design and scores on the EDCT (our test of conceptual knowledge), suggesting that our results here may apply to a broad range of students. Anecdotally, we have subsequently heard from instructors who feel their students did better on other experimental design tasks in their classes after having completed the UED tutorial, supporting this conclusion. The variation between classes may indicate that different populations would benefit from different amounts of feedback and constraint. As we argue elsewhere (Meir, 2022), changing the level of constraint in the learning environment might be a particularly powerful lever to adjust activities to maximize student learning for different populations.

Performance-based assessment is important for complex knowledge and skills

We have a few lines of evidence of a disconnect between students' declarative/conceptual knowledge of experimental design (i.e., defining terms and explaining concepts) and their procedural knowledge (i.e., designing an experiment). For example, there was a relatively high baseline performance on both the EDCT and interview probing questions, and relatively little change in performance on those assessments from pre to post (EDCT – Figure 3; interview questions – unpublished data). On the other hand, we have solid evidence of students improving in their procedural knowledge, as assessed by the MV-EDAT (Figures 4 and 5) after going through the process of designing and running an experiment (as in Section 2 of UED). We also have evidence that immediate feedback, in particular, improved their procedural knowledge and experimental design skills, as measured by the in-tutorial experimental designs (Figure 7). Overall, we saw more change in experimental design skills than in declarative and conceptual knowledge. It is also noteworthy that the two performance-based assessments (the MV-EDAT and the in-tutorial experimental designs) had some correlation, but we saw low to no correlation between the EDCT and either performance-based assessment.

This is not to suggest that teaching declarative and conceptual knowledge is unnecessary for developing higher-level skills. Indeed, we learned from earlier iterations of experimental design activities that without establishing a common baseline of vocabulary and a review of basic conceptual principles, students were not always able to benefit from feedback that relied on this declarative and conceptual knowledge. But, if improving performance is the goal, performance-based activities are important for building complex skills, and procedural knowledge is best assessed with a performance-based assessment such as the in-tutorial experiment or the MV-EDAT (or other versions of the EDAT).

There are numerous studies showing highly-constrained assessments such as multiple choice-based tests miss aspects of understanding and skills that less constrained assessments capture (e.g., Nehm *et al.*, 2012; Beggrow *et al.*, 2014; Hubbard *et al.*, 2017; Uhl *et al.*, 2021). Designing performance-based assessments with intermediate degrees of constraint may have benefits. Asking students to complete tasks with some constraints, such as the experimental design tasks in UED, may help gauge student skill level, and help focus in on exactly where a student is confused in ways that higher constraint assessments, and potentially even low-constraint assessments, cannot, while also allowing those assessments to be autoscored (Hubbard *et al.*, 2017; Meir, 2022).

Study limitations leave open other interpretations

Our own experimental design includes some inherent limitations, so the conclusions we reach above come with caveats.

While we validated the EDCT in several ways, this study was the first time it was used and there is a reasonable chance that it simply is not sensitive enough to distinguish large changes in understanding in the samples in our study. The Wright map, for instance, indicates that many items on the EDCT were easy for the samples we studied.

This study was also the first time that our revised version of the EDAT (the MV-EDAT) was used. We designed the MV-EDAT to fit our assessment needs, by creating prompts with nonhuman contexts and also deliberately suggesting more than one independent variable that could be tested, in order to assess how students deal with the realistic scenario of designing experiments with multiple putative causative independent variables. We also decided to implement the MV-EDAT with accompanying probing questions because we wanted to assess their declarative and conceptual understanding of concepts, which may not be apparent from what they volunteer on paper if they leave key concepts out of their experimental design description (e.g., what variables they would hold constant between treatments). Based on our results, we can recommend the use of the MV-EDAT prompts we designed in contexts where they might be useful (e.g., in a class focused on nonhuman rather than human biology, and when particularly interested in how students face the challenge of designing experiments with multiple independent variables). However, we did not really see an added benefit to pairing the MV-EDAT with interviews designed to probe student declarative and conceptual knowledge. This level of knowledge is well suited for constrained choice tests, like the EDCT we developed. When implementing any version of the EDAT (e.g., the original EDAT, the Expanded EDAT, or our MV-EDAT), researchers or instructors should understand that procedural knowledge is what is being tested. Our experience with the interviews shows that students not including elements in their experimental design do not mean they cannot identify or even define or explain that element. Our experience just reinforces the importance of identifying what level of knowledge you are interested in assessing and choosing the appropriate assessment for that level.

We also acknowledge that many of our conclusions comparing among treatments may be limited because they are based on an assessment task within the activity itself, rather than the more independent assessment MV-EDAT. Given the consistency on the within-tutorial assessments between the Individual and Split-Class samples, we think it likely that were we able to devote the resources to complete the full interview protocol with a larger number of students, we would have seen the same results in the MV-EDAT based on nonsignificant trends in that data. It is certainly possible, however, that feedback only affected students' ability to complete the experimental design tasks within UED, and did not have a differential effect on their ability to transfer that learning to the other context represented in the MV-EDAT. Supporting our interpretation, though is that the element scored most similarly between the in-tutorial exercises and the MV-EDAT (Replication) was correlated between the two.

CONCLUSIONS

While there is no doubt that active learning approaches are critical for mastering core scientific skills and knowledge, the phrases "active learning," "student-centered teaching," and other similar language encompass a broad range of activities. To determine which approaches within that range are most effective in different situations requires experiments that test alternatives of how to design those activities (Freeman *et al.*, 2014). Here we show that two key axes upon which learning activities can vary, feedback and constraint, are both likely to be important in maximizing learning of a core skill in biological science, although for different reasons. We show that immediate, specific feedback is highly effective for helping students learn. Our data suggests that some variation in constraint, at least within the intermediate range, may not have a large direct effect on learning. But because constraint allows feedback and has other indirect effects, degree of constraint is useful to consider as a way of maximizing learning through other avenues. While our research focused on experimental design skills, we suggest these results may also be applicable to the teaching of other skills of similar complexity.

ACKNOWLEDGMENTS

This multifaceted study would not have been possible without help from many people on and off the project team. As part of the team at various times, we greatly appreciate the intellectual contributions made by Jody Clarke-Midura, Jenna Conversano, and Eric Klopfer, administrative help from Evelyne Tschibelu and Carol Burke, and the many people at SimBiotic Software involved in putting together the special versions of UED used here including Steve Allison-Bunell, Eric Harris, Josh Quick, Derek Stal, Eleanor Steinberg, and Jennifer Wallner. We also thank our advisory board of Ross Nehm, Ryan Baker, and Kathryn Perez. We thank the editor for a heroic statistical review and two anonymous reviewers who provided feedback that greatly improved the manuscript. Outside the team we extend a special thanks to the many instructors and students who took part in the study. This material is based upon work supported in part by the National Science Foundation under Grant No.-1227245 and while serving at the National Science Foundation.

REFERENCES

Abdullah, C., Parris, J., Lie, R., Guzdar, A., & Tour, E. (2015). Critical analysis of primary literature in a master's-level class: Effects on self-efficacy and science-process skills. *CBE—Life Sciences Education*, *14*(3), ar34. https://doi.org/10.1187/cbe.14-10-0180

Abraham, J. K., Meir, E., Perry, J., Herron, J. C., Maruca, S., & Stal, D. (2009). Addressing undergraduate student misconceptions about natural selection with an interactive simulated laboratory. *Evolution: Education and Outreach*, *2*(3), 393–404. https://doi.org/10.1007/s12052-009-0142-3

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC: AAAS.

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How learning works: Seven research-based principles for smart teaching*. San Francisco, CA: Jossey-Bass.

Andrews, T. M., Leonard, M. J., Colgrove, C. A., & Kalinowski, S. T. (2011). Active learning not associated with student learning in a random sample of college biology courses. *CBE—Life Sciences Education*, *10*(4), 394–405. https://doi.org/10.1187/cbe.11-07-0061

Arthurs, L. A., & Kreager, B. Z. (2017). An integrative review of in-class activities that enable active learning in college science classroom settings, *International Journal of Science Education*, *39*(15), 2073–2091, https://doi.org/10.1080/09500693.2017.1363925

Baker, R. J. D. (2011). Gaming the system: A retrospective look. *Philipp Comput J*, *6*(2011), 9–13.

Behar-Horenstein, L. S., & Niu, L. (2011). Teaching critical thinking skills in higher education: A review of the literature. *J. College Teaching and Learning*, *8*(2), 25–42. https://doi.org/10.19030/tlc.v8i2.3554

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? *Journal of Science Education and Technology*, *23*, 160–182. https://doi.org/10.1007/s10956-013-9461-9

Blanchard, M. R., Southerland, S. A., Osborne, J. W., Sampson, V. D., Annetta, L. A., & Granger, E. M. (2010). Is inquiry possible in light of accountability? A quantitative comparison of the relative effectiveness of guided inquiry and verification laboratory instruction. *Science Education*, *94*, 577–616.

Bodine, E. N., Panoff, R. M., Voit, E. O., & Weisstein, A. E. (2020). Agent-based modeling and simulation in mathematics and biology education. *Bulletin of Mathematical Biology*, *82*, 101. https://doi.org/10.1007/s11538-020-00778-z

Brooks, C., Carroll, A., Gillies, R. M., & Hattie, J. (2019). A matrix of feedback for learning. *Australian Journal of Teacher Education*, *44*(4), 14–32. http://dx.doi.org/10.14221/ajte.2018v44n4.2

Brownell, S. E., Wenderoth, M. P., Theobald, R., Okoroafor, N., Koval, M., Freeman, S., ... & Crowe, A. J. (2014). How students think about experimental design: novel conceptions revealed by in-class activities. *BioScience*, *64*(2), 125–137. https://doi.org/10.1093/biosci/bit016

Buck, L. B., Bretz, S. L., & Towns, M. H. (2008). Characterizing the level of inquiry in the undergraduate laboratory. *J. College Science Teaching*, *38*, 52–58.

Chernikova, O., Heitzmann, N., Stadler, M., Holtzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research*, *90*(4), 499–541. https://doi.org/10.3102/0034654320933544

Clarke-Midura, J., Pope, D. S., Maruca, S., Abraham, J. K., & Meir, E. (2018). Iterative design of a simulation-based module for teaching evolution by natural selection. *Evolution Education & Outreach*, *11*(4), 1–17. https://doi.org/10.1186/s12052-018-0078-6

Colburn, A. (2000). An inquiry primer. *Science Scope*, *23*(6), 42–44.

Craig, A. R., & Fisher, W. W. (2019). Randomization tests as alternative analysis methods for behavior-analytic data. *Journal of the Experimental Analysis of Behavior*, *111*(2), 309–328. https://doi.org/10.1002/jeab.500

Dasgupta, A. P., Anderson, T. R., & Paleaz, N. (2017). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE—Life Sciences Education*, *13*(2), 265–284. https://doi.org/10.1187/cbe.13-09-0192

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2017). Development of the biological experimental design concept inventory (BEDCI). *CBE—Life Sciences Education*, *13*(3), 540–551. https://doi.org/10.1187/cbe.13-11-0218

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proc Natl Acad Sci USA*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Gobert, J. D., Sao Pedro, M. A., Baker, R. S. J.D., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*, *4*(1), 104–143.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521–563. https://doi.org/10.1080/10508406.2013.837391

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Hesterberg, T. C., Moore, D. S., Monaghan, S., Clipson, A., & Epstein, R. (2003). Bootstrap methods and permutation tests. In: *The Practice of Business Statistics* (pp. 16.1–16.57). New York, NY: W.H. Freeman & Co.

Hubbard, J. K., Potts, M. A., & Couch, B. A. (2017). How question types reveal student thinking: An experimental comparison of multiple–true–false and free-response formats. *CBE—Life Sciences Education*, *16*(2), ar26. https://doi.org/10.1187/cbe.16-12-0339

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Klopfer, E., Scheintaub, H., Huang, W., Wendel, D., & Roque, R. (2009). The simulation cycle: Combining games, simulations, engineering and science using StarLogo TNG. *E-Learning*, *6*(1), 71–96.

Kuhn, D., Pease, M., & Wirkala, C. (2009). Coordinating the effects of multiple variables: A skill fundamental to scientific thinking. *Journal of Experimental Child Psycholog,y*, *103*(3), 268–284. https://doi.org/10.1016/j.jecp.2009.01.009

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Front Psychol*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Magalhaes, P., Ferreira, D., Cunha, J., & Rosario, P. (2020). Online vs traditional homework: A systematic review on the benefits to students' performance. *Computers and Education*, *152*, 103869. www.sciencedirect.com/science/article/abs/pii/S0360131520300695

Maier, U., Wolf, N., & Randler, C. (2016). Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Computers & Education*, *95*(2016), 85–98. http://dx.doi.org/10.1016/j.compedu.2015.12.002

McConnell, D. A., Chapman, L., Czajka, D., Jones, J. P., Ryker, K. D., & Wiggen, J. (2017). Instructional Utility and Learning Efficacy of Common Active Learning Strategies. *Journal of Geoscience Education*, *65*, 604–625.

McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment Research & Evaluation*, *18*(2), 1–15.

Meir, E. (2022). Strategies for targeting the learning of complex skills like experimentation to different student levels: The intermediate constraint hypothesis. In Pelaez, N. J., Gardner, S. M., & Anderson, T. R. (Eds.), *Trends in Teaching Experimentation in Life Sciences* (pp. 523–545). Cham, Switzerland: Springer Nature Switzerland AG.

Meir, E., Wendel, D., Pope, D. S., Hsiao, L., Chen, D., & Kim, K. J. (2019). Are intermediate constraint question formats useful for evaluating student thinking and promoting learning in formative assessments? *Computers & Education*, *141*, 103606. https://doi.org/10.1016/j.compedu.2019.103606

Momsen, J. L., Long, T. M., Wyse, S. A., & Ebert-May, D. (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE—Life Sciences Education*, *9*, 435–440. https://doi.org/10.1187/cbe.10-01-0001

Nehm, R. H. (2019). Biology education research: Building integrative frameworks for teaching and learning about living systems. *Discip Interdscip Sci Educ Res*, *1*, 15 (2019). https://doi.org/10.1186/s43031-019-0017-6

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *American Biology Teacher*, *74*, 92–98.

Nehm, R. H., Finch, S. J., & Sbeglia, G. C. (2022). Is active learning enough? The contributions of misconception-focused instruction and active-learning dosage on student learning of evolution. *BioScience*, *72*(11), 1105–1117. https://doi.org/10.1093/biosci/biac073

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

Paas, F. G. W.C., & Van Merrienboer, J. J. G.V. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors*, *35*(4), 737–743.

Pelaez, N. J., Anderson, T. R., Gardner, S. M., Yin, Y., Abraham, J. K., Bartlett, E., ... & Stevens, M. (2017). The basic competencies of biological experimentation: Concept-skill statements. *PIBERG Instructional Innovation Materials*, Paper, 4. Retrieved from http://docs.lib.purdue.edu/pibergiim/4

Pelaez, N. J., Gardner, S. M., & Anderson, T. R. (2022). The problem with teaching experimentation: Development and use of a framework to define fundamental competencies for biological experimentation. In Pelaez, N. J., Gardner, S. M. & Anderson, T. R. (Eds.), *Trends in Teaching Experimentation in Life Sciences* (pp. 3–27). Cham, Switzerland: Springer Nature Switzerland AG.

Pope, D., Maruca, S., Palacio, J., Meir, E., & Herron, J. (2016). *Understanding Experimental Design*. SimBiotic Software, Missoula MT:. Simbio.com

Puntambekar, S., Gnesdilow, D., Dornfeld Tissenbaum, C., Narayanan, N. H., & Rebello, N. S. (2020). Supporting middle school students' science talk: A comparison of physical and virtual labs. *J Research Science Teaching*, *58*, 392–419. https://doi.org/10.1002/tea.21664

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria:. URL Retrieved from www.R-project.org/

Rutten, N., van Joolingen, W. R., & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers & Education*, *58*, 136–153. doi: 10.1016/j.compedu.2011.07.017

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment*, *4*(6), 1–46. http://files.eric.ed.gov/fulltext/EJ843857.pdf

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, *39*, 37–63 http://dx.doi.org/10.1016/j.dr.2015.12.001

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Sirum, K., & Humburg, J. (2011). The Experimental Design Ability Test (EDAT). *Bioscene*, *37*, 8–16.

Shanks, R. A., Robertson, C. L., Haygood, C. S., Herdliksa, A. M., Herdliska, H. R., & Lloyd, S. A. (2017). Measuring and advancing experimental design ability in an introductory course without altering existing lab curriculum. *J Microbiology & Biology Education*, *18*(1), 1–8. https://doi.org/ 10.1128/jmbe.v18i1.1194

Sweller, J., Kirschner, P. A., & Clark, R. E. (2007). Why minimally guided teaching techniques do not work: A reply to commentaries. *Educational Psychologist*, *42*, 115–121.

Theobald, R., & Freeman, S. (2014). Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. *CBE—Life Sciences Education*, *13*, 41–48. https://doi.org/10.1187/cbe-13-07-0136

Uhl, J. D., Sripathi, K. N., Meir, E., Merrill, J., Urban-Lurain, M., & Haudek, K. C. (2021). Automated writing assessments measure undergraduate learning after completion of a computer-based cellular respiration tutorial. *CBE—Life Sciences Education*, *20*(3), ar33. https://doi.org/10.1187/cbe.20-06-0122

Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., & Reece, J. B. (2017). *Mastering Biology*. New York, NY: Pearson Education Inc.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H.M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Education Research*, *85*(4), 1–37. https://doi.org/10.3102/0034654314564881

Wang, Z., Gong, S. Y., Xu, S., & Hu, X. E. (2019). Elaborated feedback and learning: Examining cognitive and motivational influences. *Computers & Education*, *136*(2019), 130–140. https://doi.org/10.1016/j.compedu.2019.04.003

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Woolley, J. S., Deal, A. M., Green, J., Hathenbruck, F., Kurtz, S. A., Park, T. K. H., ... & Jensen, J. L. (2018). Undergraduate students demonstrate common false scientific reasoning strategies. *Thinking Skills and Creativity*, *27*, 101–113. https://doi.org/10.1016/j.tsc.2017.12.004

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers and Education*, *143*(2020), 103668. https://doi.org/10.1016/j.compedu.2019.103668