Supplemental Material

CBE—Life Sciences Education

Meir *et al*.

Supplementary Materials

Designing activities to teach higher-order skills: How feedback and constraint affect learning of experimental design

Eli Meir, Denise Pope, Joel Abraham, Kerry J Kim, Susan Maruca, and Jennifer Palacio

A. Learning Outcomes for Understanding Experimental Design

Prior to writing Understanding Experimental Design, we composed a set of 17 learning outcomes we intended to target in the tutorial. We based these learning outcomes on a literature review, well summarized by the ACE-BIO Network (see Pelaez et al. 2017 reference), as well as on previous research we conducted as part of this same project. The assessments used in this study, EDCT and MV-EDAT, were written to assess these learning outcomes.

Table S.1. Learning outcomes targeted by Understanding Experimental Design. The "Assessed by" column shows where we measured student understanding of each concept in this study.

Learning outcomes for *Understanding Experimental Design* (17 total)	**Assessed by**
Independent variable	
A. Given experiment, identify IV	EDCT Q1
B. Given scenario & hypothesis, choose appropriate IV	EDCT Q8
Learning outcomes for *Understanding Experimental Design*, continued	**Assessed by**
Dependent variable	
C. Given experiment, identify DV	EDCT Q2
D. Given scenario & hypothesis, choose appropriate DV	EDCT Q9
E. In design/execution of an experiment, record data with appropriate DV	MV-EDAT
Control and experimental treatments	
F. Given experiment, identify control and experimental treatments	EDCT Q4
Potentially confounding variables	
G. Given experiment, identify PCVs that are (or should be) held constant	EDCT Q12
Systematic variation (encompasses concepts of: IV, control & experimental treatments, holding constant PCVs)	
H. Given scenario & hypothesis, choose appropriate systematic variation, with appropriate IV and control & experimental groups, and holding constant PCVs	EDCT Q6
I. Explain why systematic variation (varying only a single variable between treatments) is important for inferring causality	EDCT Q7
J. In design/execution of an experiment, create systematic variation, with appropriate IV and control & experimental groups, and holding constant PCVs	MV-EDAT

Replication/Experimental units	
K. Given experiment, identify replicates/experimental units	EDCT Q3
L. Given scenario & hypothesis, choose appropriate replication	EDCT Q10
M. Explain why replication is important for reducing uncertainty/increasing scope of inference	EDCT Q11
N. In design/execution of an experiment, create replicates of both control and experimental treatments	MV-EDAT
Interpreting data/drawing conclusions	
O. Given scenario, hypothesis & data, infer if data support or reject hypothesis	EDCT Q5
P. In execution of an experiment, infer if data support or reject hypothesis	Not assessed
Q. In execution of an experiment, relate conclusion back to real world issue	MV-EDAT

B. Screening Survey

This assessment included seven multiple-choice questions about the design of biology experiments (five from the Test of Scientific Literacy Skills by Gormally et al, 2012, and two from D'Costa & Schlueter, 2013), and two questions about natural selection to assess general biology knowledge (one multiple-choice question from Bishop & Anderson, 1990, and one open-response question from the ACORNS instrument by Nehm et al, 2012). We used the open-response question primarily to screen out students who did not answer, since we wanted students willing to write out and verbally respond to open-ended questions during the in-person experimental design task and follow-up interviews. Based on scores to the eight multiple-choice questions, we split students into three bins - high, medium, and low performing. Approximately half of the students scored 75% correct on the multiple-choice questions and were binned into the "medium" category, so the test did not discriminate among students as well as we would have liked.

C. Experimental Design Concepts Test

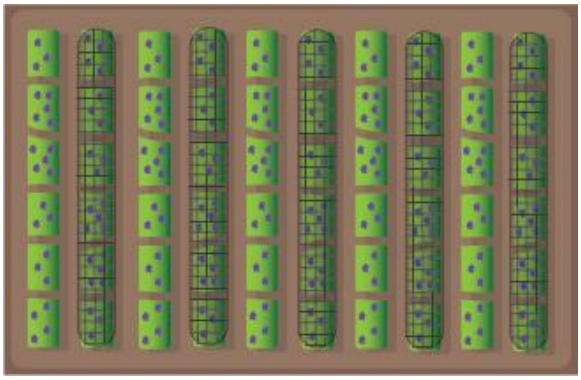
As described in the body of the paper, we wrote the Experimental Design Concepts Test to capture many of the learning outcomes shown in Table S.1, specifically those that were targeted in UED Section 1, which focused on building vocabulary and understanding of basic concepts in experimental design.

C.1 The EDCT Assessment

Use the following information to answer the first 5 questions

A farmer who grows blackberries has found his harvest to be lower than usual. He suspects that some animals are eating the berries before he can pick them. The farmer thinks it may be either birds

(landing on the bushes to feast on berries) or rodents (climbing the bushes from below to munch on berries). He decides to first test the hypothesis that birds are eating his berries, and consults with a local expert on how best to prevent the birds from eating berries. He has eight fields with 10 rows of berry bushes in each field. In one of his fields (depicted below, as viewed from above), he covers half of the rows with nets that will prevent birds from landing on the bushes, and leaves the other half uncovered. Over the course of 8 weeks, he counts the berries harvested from each row of bushes in the test field.



1. What is the independent (treatment) variable in the farmer's experiment?
    a. Presence/absence of birds
    b. Presence/absence of nets
    c. Number of rows
    d. Number of berries harvested

2. What is the dependent (response) variable in his experiment?
    a. Presence/absence of birds
    b. Presence/absence of nets
    c. Number of rows
    d. Number of berries harvested

3. What are the experimental units in his experiment?
    a. The blackberries
    b. The blackberry bushes
    c. The rows of bushes
    d. The fields

4. What is the control group in his experiment?
    a. The uncovered rows
    b. The covered rows
    c. The other 7 fields
    d. There is no control group

5. After 8 weeks, the farmer found no difference between groups, and the berry harvest was still lower than previous years. What can he conclude about his hypothesis?
    a. Rodents are eating the berries.
    b. Birds are eating the berries.
    c. Birds are not eating the berries.
    d. The results are inconclusive.

Use the following information to answer the next 3 questions

A group of researchers is planning an experiment to test the hypothesis that sleep deprivation decreases memory retention. All participants will have a practice session in which they are given strings of words to memorize. That night they will all sleep in a monitored lab for a specified time depending on their group. Then each participant will take a test to see how many words they remember from the practice session. The researchers need to decide what conditions they should change for the participants in their experimental group(s). The table below shows the conditions for the control group and three possible experimental groups.

	Control Group	Experimental Group 1	Experimental Group 2	Experimental Group 3
Number of hours of sleep	8 hours	8 hours	4 hours	4 hours
Days between practice and testing	1 day	1 day	1 day	3 days
Length of practice session	30 minutes	45 minutes	30 minutes	30 minutes

6. Which experimental group(s) should the researchers use to compare to the control group in order to test their hypothesis?
    a. Experimental group 1
    b. Experimental group 2
    c. Experimental group 3
    d. All 3 experimental groups

7. A good experiment MUST include …
    a. groups that differ only in the hypothesized causal factor(s).
    b. groups that will tell you if a factor other than the hypothesized causal factor(s) may be influencing the result.
    c. groups that differ in multiple factors.
    d. groups that will allow you to make as many comparisons as possible.

8. Many factors can potentially affect the outcome of an experiment, but the researchers are unlikely to be able to hold constant all possible factors. Which of the following factors is the most important for the researchers to hold constant in their experiment in order to test their hypothesis?
    a. Having all participants sleep the same number of hours
    b. Having all participants tested on the same day of the week
    c. Keeping the temperature of the testing rooms the same as the sleeping rooms
    d. Having all participants in the same age range (20-30 years)

Use the following information to answer the next 2 questions

Advertisements for a new energy drink claim that it helps people concentrate, and you are designing an experiment to test the hypothesis that the drink improves concentration. A large group of study participants are going to come into your lab and take a concentration test. Now you must choose independent and dependent variables for your experiment.

9. What would be an appropriate independent (treatment) variable for your experiment?
    a. How much of the energy drink participants choose to drink right before the test
    b. How much of the energy drink participants report drinking in the last month
    c. How well participants can concentrate during the test
    d. How much of the energy drink participants are told to drink right before the test

10. Of the following possible dependent (response) variables, which is the most appropriate for your experiment?
    a. Participants' self-reported ability to concentrate on the concentration test
    b. Participants' score on the concentration test
    c. Participants' heart rate during the concentration test
    d. Participants' ability to stay awake during the concentration test

---

Use the following information to answer the next question

Many gardeners believe that certain edible herbs, like oregano, sage, and rosemary, taste better if deprived of water and nutrients. You decide to test the hypothesis that the frequency of watering influences the flavor of sage. You grow two sage plants from seed, side by side in separate pots. You water one every other day (as a control treatment) and the other only once a week (as an experimental treatment). Then you invite 20 friends over for a blind taste test, giving them each two samples of sage leaves, marked only "A" and "B". Your friends rate how flavorful each sample is on a scale of 1-5. You compare the average scores for each plant.

11. As with any experiment, your setup has limitations. Which of the following ideas for increasing replication will most improve your ability to test the hypothesis that watering frequency affects the taste of sage?
    a. Increase the number of people you ask to do the blind taste test.
    b. Increase the number of leaf samples taken from each plant.
    c. Increase the number of potted sage plants in the control and experimental groups.
    d. Do the same test on rosemary and oregano, in addition to sage.

For the next 3 questions:

Would a biologist agree or disagree with the following explanations for why replication is important to consider when designing an experiment?

12. Replication reduces the chance that an uncommon result will lead you to an incorrect conclusion.
    a. Agree
    b. Disagree

13. Replication makes it possible to compare the experimental treatment to the control treatment.
    a. Agree
    b. Disagree

14. Replication increases the certainty that your results apply more widely and not just to specific cases.
    a. Agree
    b. Disagree

C.2 Lines of evidence for EDCT assessment validity

Since the focus of this study was to assess the effectiveness of feedback and constraint on student learning of a complex higher-order skill and not to develop a new instrument, we did not conduct full-scale validation efforts of the Experimental Design Ability Test. However, we did collect several lines of validity evidence. While our validation efforts are less extensive than those seen in the development of recent concept inventories intended for general use, the combination of evidence we collected gave us enough confidence that the EDCT (Section C.1 above) was measuring student performance on the focal learning outcomes of Section 1 of UED (section A above), which was the purpose for which we used it here.

Expert review

We asked four faculty to review the EDCT, none of whom were involved in writing the assessment. Three were university faculty who publish educational research. The fourth was a faculty member who oversees a large introductory biology class where experimental design is taught. The faculty rated each question in the EDCT for three criteria on a scale of 1 - 5 (5 being best): (1) Scientific accuracy, (2) Clarity, and (3) Relevance to learning outcome.

We separately had eight faculty with self-identified expertise perform blooming (Crowe et al, 2008) on the EDCT questions. We averaged these scores to assign a Blooms level from 1 - 6 to each question, where 1 = Remember; 2 = Understand; 3 = Apply; 4 = Analyze; 5 = Evaluate; and 6 = Create. There was a distinct split with 8 questions rated as lower level (Blooms score < 3) and 6 questions rated as higher-level (Blooms score > 3).

Table S.2 Reviews of EDCT items. Scientific Accuracy, Clarity, and Relevance were all scored by 4 faculty / BER researchers and the table shows average score for each item. Blooms Level was scored by 8 self-identified Blooming experts. The column shows their average score for each item; items identified as being at higher Blooms levels (>3) shaded in gray.

	Scientific Accuracy	Clarity	Relevance	Blooms Level
Q1	4.5	4.3	4.5	2.6
Q2	4.0	4.5	4.5	2.6
Q3	4.5	4.3	4.5	2.8
Q4	4.5	4.5	4.5	2.6
Q5	4.3	3.3	4.5	3.9
Q6	4.5	4.5	4.5	4.1
Q7	4.5	4.3	4.0	1.8
Q8	3.8	4.0	4.5	4.4
Q9	4.0	4.0	4.5	4.1
Q10	4.5	4.5	4.5	4.0
Q11	4.5	4.5	4.5	4.3
Q12	4.5	4.5	4.5	2.5
Q13	4.0	3.0	3.5	2.4
Q14	4.0	4.0	4.0	2.5

Student interviews

After developing the close-to-final version of EDCT, we interviewed 8 students recruited from colleges in Massachusetts and Utah using a think-aloud protocol, where we asked them to express their thinking as they selected their answers for each question and the interviewer probed students on any confusions they seemed to have. In between the interviews, we revised questions where confusions seemed related to wording rather than to naive thinking about experimental design.

With the final version of the EDCT, we interviewed an additional 4 students with the same think-aloud protocol. In general, these students appeared to understand the question stems and answers, and in most cases their verbalized thoughts indicated that the answers they chose reflected the understanding of experimental design we were trying to capture.

Internal reliability

We estimated internal reliability of the EDCT with data pooled across two introductory biology courses from different western US public 4-year institutions (n = 124). The Cronbach's alpha was 0.64, which is a little lower than desired for a college assessment but within the range of some other published tests on biology concepts (e.g. Perez et al, 2013). Some coverage of experimentation in biology had already occurred in both courses prior to the assessment.

C.3 Rasch Analysis of EDCT

We conducted Rasch analysis (Boone 2016) to assess the performance of the EDCT using R (R Core Team, 2018) and the 'TAM' and 'WrightMap' packages. We calculated Item Difficulty and Item Fit to evaluate the suitability of each item and the performance of the test as a whole. One data set (n = 165) analyzed was collected in an introductory biology course at a comprehensive master's granting institution in the western US (the "Split-Class Sample", described in Section 2.4.2), in which we administered the EDCT as a pre- and post-instruction assessment. We also analyzed post-instructional EDCT data (n = 1292) collected from an additional 27 courses, ranging from introductory to upper-division biology (the "Larger-Scale Sample," described in Section 2.4.3).

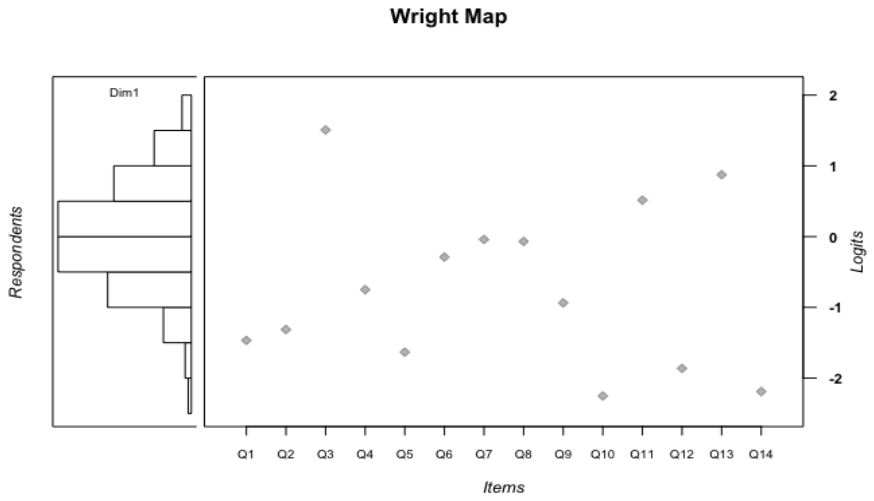
All items in the EDCT were within acceptable ranges with regards to Infit and Outfit MNSQ in the pre- and post-instruction implementations (Supplementary Materials, Section B3). Item difficulty was low enough on three items (10, 12, 14), that they provided little information about student performance. However, the EDCT as a whole has a reasonably wide enough range of item difficulties to discriminate across students.

Split-Class Sample (Pre- and Post-Instruction)

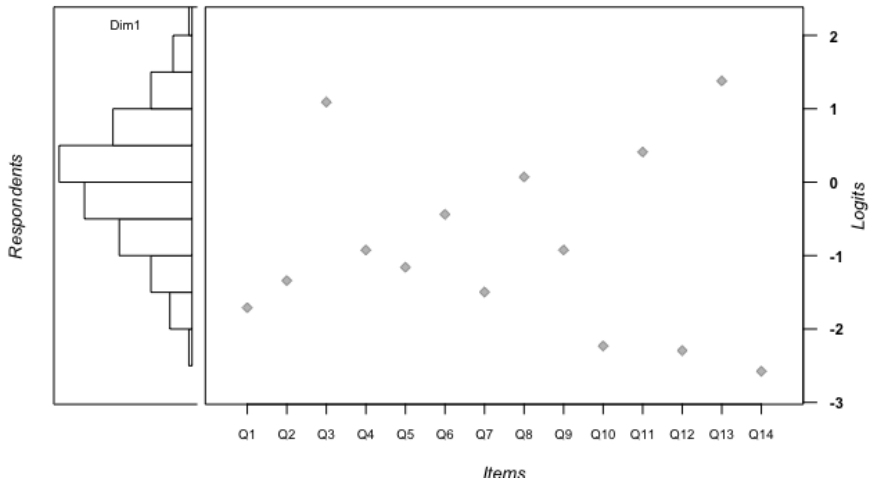
Item information. The low difficulty of items 10, 12, and 14 mean they provide little information to differentiate student performance. Note that item 10 was one identified as being at a higher Blooms level, so item difficulty does not necessarily correspond with Blooms level.

Item	Pre Difficulty	Post Difficulty	Pre Outfit	Post Outfit	Pre Infit	Post Infit
Q1	-1.467	-1.710	0.930	0.903	0.942	0.940
Q2	-1.314	-1.341	0.989	0.952	0.978	0.959
Q3	1.507	1.091	1.164	1.147	1.060	1.077
Q4	-0.750	-0.924	1.030	0.974	1.011	0.976
Q5	-1.633	-1.159	1.057	1.082	1.028	1.068
Q6	-0.289	-0.438	0.941	0.869	0.950	0.895
Q7	-0.041	-1.497	0.964	0.863	0.966	0.935
Q8	-0.068	0.071	1.055	1.071	1.053	1.049
Q9	-0.937	-0.924	0.923	0.908	0.959	0.937
Q10*	-2.252	-2.231	0.845	0.9723	0.9611	0.988
Q11	0.515	0.411	1.027	1.070	1.020	1.042
Q12*	-1.863	-2.293	0.971	0.940	0.985	0.986
Q13	0.874	1.379	1.016	1.195	1.000	1.080
Q14*	-2.190	-2.576	1.161	1.289	1.072	1.082

Pre-Test



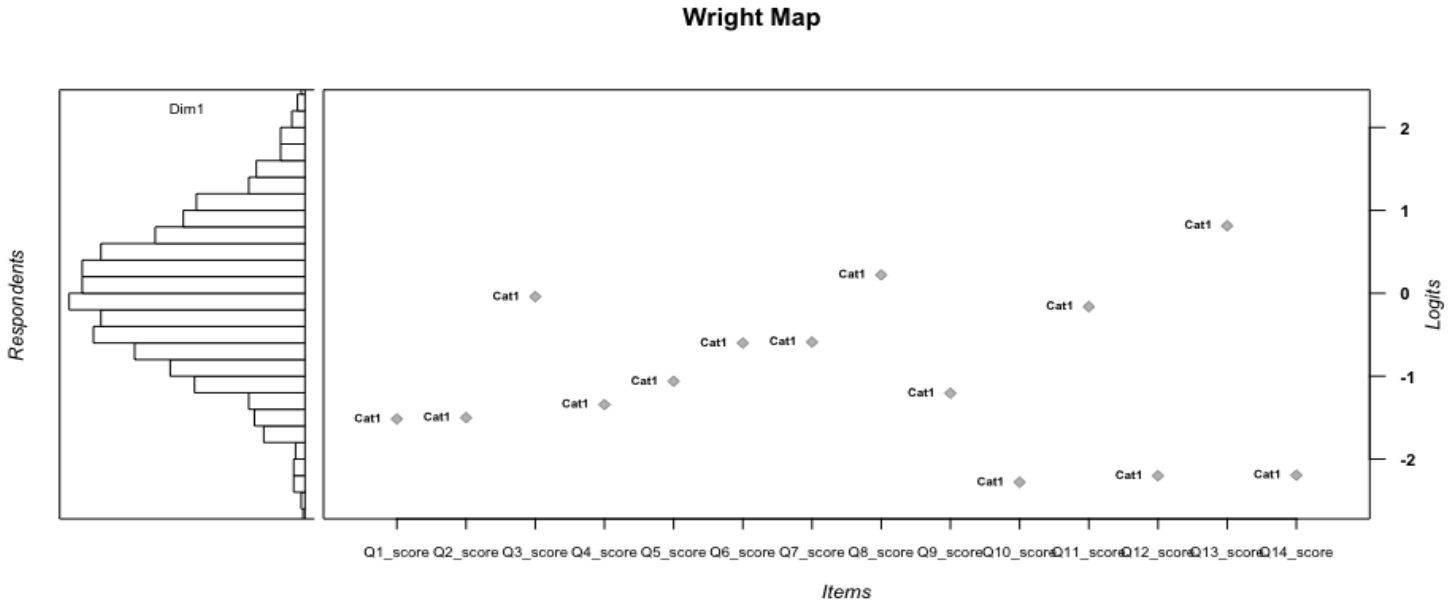
Post Test



Large-Scale Sample (Post-instruction)

Item Information. The low difficulty of items 10, 12, and 14 mean they provide little information to differentiate student performance.

Item	Difficulty	Outfit	Infit
Q1	-1.516	0.918	0.956
Q2	-1.501	0.952	0.968
Q3	-0.041	1.001	1.001
Q4	-1.343	0.936	0.968
Q5	-1.061	1.0478	1.039
Q6	-0.600	0.906	0.928
Q7	-0.589	0.897	0.919
Q8	0.221	1.149	1.110
Q9	-1.205	1.041	1.026
Q10*	-2.279	0.919	0.969
Q11	-0.161	1.059	1.046
Q12*	-2.202	0.883	0.977
Q13	0.814	1.007	1.008
Q14*	-2.194	1.160	1.066



D. Multiple Variable Experimental Design Ability Test (MV-EDAT)

For the study, we wanted data on what learning students could transfer from the UED tutorial to a new, independent context. For this purpose, we looked for a pre / post assessment of experimental design procedural knowledge that was open-ended, could capture many of the skills that UED was designed to teach, and was independent of the tutorial. We started with the Experimental Design Ability Test (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). We created revised prompts to better fit our context, with two specific differences from the original EDAT prompts. First, since our tutorial focused on a biological example of a hypothetical animal species, we wanted to use non-human animals in our prompt, so we created examples using lizards and fish (deliberating choosing non-mammalian examples that students are less familiar with). This choice eliminates some of the elements that the EDAT tested for, such as the necessity of placebo controls, which weren't relevant for our non-human context. Second, since one of the common mistakes when learning experimental design is attempting to test too many independent variables in a single experiment, and UED was designed to teach that concept and test students' ability to apply it when designing experiments, we wanted to create examples that suggested multiple possible independent variables that could be tested.

We thus designed our own two parallel prompts which were modeled after the EDAT but included modifications to address those two requirements. We call these the Multiple Variable EDAT (MV-EDAT), and the two versions are called "Lizard" and "Fish" prompts after the species used in the prompt (section D1 below). Each student received one version as a pre-assessment and the other version as a post-assessment, randomly assigned. They answered the prompts by drawing and/or writing on the paper that included the prompt.

In addition to the differences in the prompts we created, we also implemented the MV-EDAT differently than the original EDAT and Expanded EDAT. We learned from pilot tests using the Expanded EDAT that if students failed to include some aspect of experimental design (e.g. sample size, or what they would measure) in their written description of their experiment, we were uncertain whether they did not understand the concept or had an implicit understanding of it but neglected to include it explicitly in their description. To more completely document their declarative and conceptual knowledge, we decided to pair the pen & paper experimental design task with a follow-up semi-structured interview. 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).

D.1 The MV-EDAT Assessment

The MV-EDAT Fish Prompt

DESIGN AN EXPERIMENT

Background information:

You work for an aquarium that is part of a breeding program for an endangered fish species. Your aquarium sends fish eggs to other aquariums to increase the population. In the wild, the species ranges in color from a deep blue to a bright white. Each female lays eggs from April through September and produces 20 eggs per month on average during that time. However, in your aquarium's breeding program, the females are producing an average of only 4 eggs per month. The initial colors of your breeding fish were similar to the ratios seen in the wild, but now more of them are white. You have been asked to investigate how to increase the egg-laying rate of the females in your breeding program.

In your breeding program, fish pairs are kept in separate tanks to mimic the territories they establish in the wild. They are fed an artificial fish food instead of the diet of shrimp and algae that this species eats in the wild. The lights in the fish tanks are on for 12 hours per day, whereas in their wild habitat the natural daylight varies between 12 and 16 hours per day over the breeding season. The temperature in the fish tanks is kept constant at 15 degrees Celsius whereas in the wild it can range between 10 – 20 degrees. The aquarium is looking for the most cost-effective way to increase the fish's egg-laying rate.

You have been asked to do an experiment to determine which environmental factor or factors should be changed to increase the average egg-laying rate for this endangered fish species in your aquarium.

Instructions:

Take 10-15 minutes to plan out your experiment so you can explain it to the interviewer.

- Write out the hypothesis you are testing with your experiment.
- Draw out your experiment and describe all of the important elements of the experiment.

The MV-EDAT Lizard Prompt

DESIGN AN EXPERIMENT

Background information:

You work for a tropical eco-park that is part of a breeding program for an endangered lizard species. Your conservation team sends lizard eggs to other zoos and parks to increase the population. In the wild, the species ranges in color from a light green to a bright purple. Each female lays eggs from October through May and produces 15 eggs per month on average during that time. However, in your park's breeding program, the females are producing an average of only 3 eggs per month. The initial colors of your breeding lizards were similar to the ratios seen in the wild, but now more of them are green. You have been asked to investigate how to increase the egg-laying rate of the females in your breeding program.

In your breeding program, lizard pairs are kept in separate tanks to mimic the territories they establish in the wild. They are fed a compressed pellet food instead of the diet of flies and beetles that this species eats in the wild. Simulated tropical rains occur in the lizard tanks for 2 hours per day, whereas in their wild habitat the natural storms last between 1 and 4 hours per day. The temperature in the lizard tanks is kept constant at 24 degrees Celsius whereas in the wild it can range between 21 – 27 degrees. The park is looking for the most cost-effective way to increase the lizard's egg-laying rate.

You have been asked to do an experiment to determine which environmental factor or factors should be changed to increase the average egg-laying rate for this endangered lizard species in your park.

Instructions:

Take 10-15 minutes to plan out your experiment so you can explain it to the interviewer.

- Write out the hypothesis you are testing with your experiment.
- Draw out your experiment and describe all of the important elements of the experiment.

D.2 Administration and scoring of the MV-EDAT

Two of the authors [DP and JP] 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. One of the interviewers was randomly assigned to interview each student for the pre-assessment, and then the other would conduct the post-assessment interview. Each interviewer did the same number of pre- and post-assessment interviews. The interviewers also set students up on the computer to start the UED tutorial, but since students were assigned login IDs, which had been pre-assigned to UED versions, by another team member, the interviewers weren't aware of which tutorial version the students were completing.

We recorded audio from the interviews. The recordings were labeled with the students' login ID and transcribed using a transcription service. A team member who was not involved in scoring the interview responses assigned each transcript a code and stripped away any preliminary or wrap-up questions to ensure the transcript scorers could not identify the student being interviewed, or whether it was a pre- or post-assessment interview.

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, if 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 6 different elements: (1) Uses systematic variation, (2) Addresses hypothesis, (3) Includes replication, (4) Includes variables held constant, (5) Includes dependent variable, (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 (e.g., systematic variation of some but not all independent variables, or inadequate

control treatments, or replication of some but not all treatments), and 2 being full and correct expression of the element (e.g., fully systematic variation, replication of all treatments).

We looked at each of the six experimental elements individually, and also combined them into a total experimental design score summing all six elements (for a total possible score of 12). Using a randomization test (see below), we found no significant difference in student pre-assessment scores between the two MV-EDAT prompts (Fish or Lizard), either in individual categories (p values of 0.22 - 0.55) or the total score (p=0.39), so we consider the two prompts to be equivalent and did not test for order effects.

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 since the semi-structured 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). For example, in response to the question "what is the same between treatments/groups?" (designed to probe understanding of potentially confounding variables) responses that did not identify any similar elements were scored as 0, those that identified one reasonable potentially confounding variable were scored as 1, and those that identified two or more were scored as 2.

For all the interview scoring (experimental design elements and probing questions), two team members independently scored each element or interview response and then the two came to consensus on each score.

E. Analysis of experimental designs within UED

Our scoring of systematic variation and appropriate controls both require a bit more explanation. By systematic variation, we mean that the design must include variation among plots in at least one variable (Simploids, Plants, Herbicide or Parasites), and that variation must be systematic (i.e., only one variable changed at a time). The scoring algorithm checked for systematic variation by checking that any one plot in an experimental design must have at least one sister plot where one and only one variable has been changed. For instance, if the design has a plot with 10 Simploids, 10 Plants, Herbicide, and no Parasites, a sister plot could be one with 10 Simploids, 10 Plants, no Herbicide and no Parasites. However, a plot with 20 Simploids, 20 Plants, Herbicide and no Parasites would not be a sister plot because both the number of Simploids and the number of Plants have changed. This also means that designs that only included replicates (e.g., 6 plots with 10 Simploids, 10 Plants, Herbicide, and no Parasites) would be scored as 0 for systematic variation.

By appropriate controls, we mean that if they used one of the putative causal variables (Herbicide or Parasites), they had to include a negative control of that variable, i.e., a plot which did not include that causal variable. So, for instance, if they added Herbicide to all plots, they were given a "Control" score of 0.

For the Intermediate Constraint treatments (ICWF and ICNF) we were able to determine all the above algorithmically. Since the LCNF treatment allowed for many more orders of magnitude of possible combinations of variables, we could not score the designs automatically, so for the LCNF treatment we determined the first three scores above manually following a rubric that matched the algorithmic determination for the other treatments as closely as possible, but with a 20% tolerance for different values between plots (i.e. if one plot contained 20 Simploids and another contained 24, those were considered the same to avoid lowering scores for unintended variation). Two raters scored each student design in the LCNF treatment, and reached consensus on each item. The full rubric for scoring LCNF data is available from the authors.

F. IRB Approval

The Committee on the Use of Humans as Experimental Subjects, the institutional review board at the Massachusetts Institute of Technology in Cambridge, MA, approved this study before data collection (COUHES #1206005102R005 and #01366099), and for each of the classes whose data we used, we also received approval from the IRBs of their institutions (they either chose to review and approve the study or accepted the approval of MIT COUHES).

G. References

Bishop, B.A., & Anderson, C.W. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching* 27(5), 415–427. https://doi.org/10.1002/tea.3660270503

Brownell, S.E., Wenderoth, M.P., Theobald, R., Okoroafor, N., Koval, M., Freeman, S., Walcher-Chevillet, C.L. & 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

D'Costa, A.R., & Schlueter, M.A. (2013). Scaffolded Instruction Improves Student Understanding of the Scientific Method & Experimental Design. *The American Biology Teacher* 75(1): 18-28. https://doi.org/10.1525/abt.2013.75.1.6

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a Test of Scientific Literacy Skills (TOSLS): Measuring Undergraduates' Evaluation of Scientific Information and Arguments. *CBE-Life Sciences Education* 11(4): 364-377. https://doi.org/10.1002/tea.3660270503

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.

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