USE OF SMALL-SCALE CLASSROOM EXPERIMENTS TO INFORM STATISTICS (SBI) PEDAGOGY IN TERTIARY CLASSROOMS

Beth Chance[1], Sophia Chung[1], and Nathan Tintle[2]
[1]California Polytechnic State University, San Luis Obispo, CA, USA
[2]Superior Statistical Research, Sioux Center, IA, USA
bchance@calpoly.edu

Typically research in education is done through observational studies, often focusing on end-of-course grades or exam scores. We explored the use of smaller-scale, focused experiments performed in genuine classroom settings. Six instructors across two institutions implemented a series of experiments focused on implementation decisions in teaching a simulation-based inference curriculum. We found the experiments to be feasible and very valuable in helping to understand whether, and why, some pedagogical methods are better than others. We offer suggestions for larger scale implementation.

INTRODUCTION

Although randomized experiments are the gold standard for comparing treatments (Scheaffer et al., 2007), they are not commonly used in educational studies due to ethical, authenticity, and logistical concerns (e.g., Chance & Garfield, 2001). In this paper, we explore the feasibility of carrying out small-scale classroom-based experiments within an existing introductory statistics course. Our context has been the study of "simulation-based inference" (SBI) curricula (e.g., Chance & Rossman, 2006; Cobb, 2007). Previously our research on the effectiveness of SBI curricula has focused on measures of student learning in introductory statistics courses comparing SBI to non-SBI courses using pre/post course measures of conceptual understanding. Evidence is mounting that this approach not only improves student understanding of statistical inference but also improves understanding of the investigative process as a whole (e.g., Chance et al., 2022; Tintle et al., 2011, 2018). However, are there "optimal" ways of introducing "simulation-based inference" to these students? Can we effectively use classroom-based experiments to explore different approaches to improving student understanding of these simulation models? We describe trial studies we carried out during the past two years and provide advice for larger-scale implementation.

THE SETTING

With the move to virtual learning due to the COVID-19 pandemic, it became much more feasible for us to manipulate learning modules given to individual students to read/view/engage in outside of class. In particular, many learning management systems (e.g., Canvas) make it easy for instructors to randomly assign students to different questions in an online quiz format. Although this feature was presumably developed to help guard against cheating, we saw an opportunity to explore different instructional strategies. For example, question 1 on a quiz can randomly assign students to different learning "demonstrations," and then the remaining questions can be the same for all students.

The *Introduction to Statistics Investigations* textbook (Tintle et al., 2020) uses a novel sequencing of content (spiraling through a six-step statistical investigation process) with heavy reliance on using simulations to explore *p*-values and confidence intervals starting from the first chapter, through the use of freely available student-driven online applets. These materials also emphasize active learning through use of explorations to accompany expository examples and student-focused statistical applets to add interactivity and visualizations of key concepts. Although we have seen promise in these materials, our studies have previously focused on end-of-course summative evaluations. Day-to-day implementation questions still remain as to what aspects of the curriculum are most beneficial and how to better improve students' learning trajectory through the material. For example, should we give students color-coded index cards or ask students to write "success" and "failure" on each card themselves? We were especially curious about whether we could replicate the "hands-on" activities we used in our courses pre-COVID-19 when carrying out simulations for the first time (e.g., coin tossing) during the pandemic-induced online learning environment.

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org ©2022 ISI/IASE

PARTICIPANTS

Over the past two years, six instructors across two universities in the United States using the *ISI* text in an algebra-based introductory statistics course agreed to participate. Each course section enrolled approximately 35 students. The assignments were integrated into existing materials (e.g., existing reading quizzes or pre-lab assignments) as much as possible. We wanted the assignments to be "medium stakes," so that students would put forth their best effort but not feel unduly punished for not performing well. In particular, the assignments were more of a "readiness quiz" format, preceding more formal instruction on the topic. Students were asked to complete a consent form at the beginning of the course (monitored by a student research assistant). In most cases, students were individually randomized to the different assignments; in one case students were randomized at the section level. Students were usually given full credit on the assignments after completion, but often did not know in advance that they were doing anything differently from their peers on the same assignment.

INTERVENTIONS

Most of the trial interventions focused on introducing students to a simulation model and how the model could be used to make inferences about a genuine student result, with the quiz questions focused on assessing student understanding of the simulation process (Table 1). For example, Study 1 had students either flip a coin to mimic randomness in a binary variable (can dogs understand human cues by correctly selecting one of two indicated objects), and then move to the One Proportion applet to carry out many such repetitions to build a null distribution. During remote teaching, we wondered whether we could use Jamboard (a "digital whiteboard," jamboard.google.com) to replicate the tactile experience and how that would compare to moving directly to the applet: some students were asked to place a sticky-note with their name above their sample result and return to the website later to see how the distribution of statistics was forming before answering the quiz questions; other students were asked to use the applet to carry out 20 repetitions of the simulation and view the resulting dotplot.

Table 1. Mini-experiments conducted 2020–2022

Study 1	*One proportion inference*
	Version A: Ask students to flip coins and record (and review) results in Jamboard
	Version B: Ask students to use applet to flip coins and create dotplot with 20 results
Study 2	*Comparing two proportions*
	Version A: Ask students to prepare and shuffle cards
	Version B: PowerPoint animation of shuffling process
	Version C: Instructor-led demonstration video of applet
	Version D: Student use of applet
Study 3	*Parameter vs. Statistic*
	Version A (Definitions): Assigned reading from text
	Version B (Module): Students work through online module, matching terms to definitions
	Version C (Video): Video of instructor explaining definitions
Study 4	*Comparing two proportions*
	Version A (Video1): Video demonstration of applet
	Version B (Video2): Voice-over PowerPoint animation of shuffling process
	Version C (zyBooks): zyBooks section with animation and instant feedback multiple choice

Similarly, another foundational simulation model emphasized in simulation-based inference curricula such as *ISI* focuses on having students shuffle index cards to mimic random assignment and build a null distribution comparing two treatment proportions. Previously, students carried out this card shuffling process in class. Questions we want to explore include whether an instructor-led video can also be used to help students understand this simulation model, does it matter whether the video is stand-alone or produced by the instructor, and how does this compare to asking students to proceed directly to the applet (Studies 2 and 4).

Study 4 had some of the same versions as Study 2 but included a zyBooks option. zyBooks (www.zyBooks.com) produces interactive textbooks and the zyBooks-adaptation of the *ISI* text is now being tested. Each zyBooks chapter consists of several subsections, and each subsection involves a

short "say" component, a "show" component (an animation of the key ideas), and an "ask" component (a series of multiple-choice questions with instant feedback for correct and incorrect answers). Students assigned to this demonstration were expected to answer these questions in the zyBooks module before continuing to the quiz questions. For Study 4, an individual subsection was shared with the students through an embedded Canvas link. Study 4 also compared two videos, one from the ISI e-book and one from an ISI instructor (neither were the course instructors for this study). Both videos focus on the building blocks of a randomization test. The first uses an earlier version of the Two Proportions applet and the second focused on a PowerPoint animation of individuals being shuffled back to the groups (same as Study 2, where some positive effects were found).

Study 3 did not focus on building understanding of a simulation model but on learning/ applying basic terminology during the first week of the course. The versions used in the study included a static reading assignment from the text, an instructor video (by the course instructor) explaining the definitions, and an online module that had students use pull-down menus to select responses with immediate feedback about the terms.

PRELIMINARY RESULTS

Across 3 terms (Winter 2021, Fall 2021, Spring 2022), two or three instructors implemented one or two of the studies at their own institutions in online and face-to-face formats. Considering the small sample sizes and the preliminary nature of these studies, we do not offer a formal analysis of these results. Chance et al. (2021) reported on Studies 1 and 2, including general observations:

- Having students "crowd source" the null distribution so each student is "one of the dots," appears to help students understand the process more than only generating a small number of repetitions with the applet.
- In the transition from the actual study to the simulation model, more care should be taken in helping students map to the study context. For example, in Study 1 student performance was not the same when the quiz questions related directly to the context and when they related to the simulation results. In Study 2, we also conjectured that some improved performance in the PowerPoint group arose from the depiction of people being shuffled into groups rather than index cards. This led us to change the depiction in the online applet from rectangles to people icons.
- Students struggled the most with mapping the "observational unit" and "variable" questions to the simulated null distribution. Perhaps these terms should be reserved for the study context, but the concept of "what is this a distribution *of*" needs to be contrasted between the study and the null distribution and continually reinforced.
- Studies 1 and 2 were in the same course and students did not exhibit complete carry over in concepts between units. We conjecture that students will benefit with more direct instruction in how the concept of a randomization distribution directly builds on what they learned in the one variable scenario. Anecdotal evidence suggests that use of the mantra "could this have happened by random chance alone?" may be helpful.
- Building a simulation model is also less intuitive for students and can benefit from direct instruction. For example, the *ISI* materials now include explicit mappings of the simulation step to the study context (e.g., "one coin toss = one choice by the dog").

Building on these observations, Studies 3 and 4 were carried out by two instructors, one face-to-face and one online, in Spring 2022. Tables 2 and 3 summarize the sample sizes, proportion correct, and average score across the demonstrations and instructors. (Question 1 in Table 3 was from a pull-down menu matching question worth 3 points.) For Study 4, students were also asked to rate their perceived effectiveness of the demonstration. Not surprisingly with the small sample sizes, not much was significant apart from some demonstration × instructor interactions, and Study 3 showed significantly lower average scores comparing the module to the instructor video for Instructor 2 (p-value = .03). We again focus on general impressions from the data as well as direct student feedback:

- Students generally performed well on initial terminology questions but continued to struggle with some of these concepts (e.g., parameter versus statistic) throughout the course. Additional analysis could track the students across the treatment groups through to other assessments in the course.

- Students in the online course, which made heavier use of videos throughout the course, may have been more successful with the video presentations than students in the face-to-face course, even when the video narrator is not the instructor, due to their familiarity and course expectations.
- Demonstrations that required more student interaction (e.g., answering zyBooks questions) appeared to have some benefits. We are still analyzing the similarity of zyBooks practice questions with the actual quiz questions to assess the degree of transfer but did find evidence that forcing more interaction during the reading process versus a voice-over PowerPoint was more effective.
- The demonstrations focused on the individual repetitions in each simulation. Many students still struggled to then "count the dots" to estimate a *p*-value, indicating that this "transfer" of knowledge (despite being similar to earlier units in the course) was still difficult for many. Students also struggle with stating a correct conclusion for an insignificant *p*-value, choosing "evidence for the null hypothesis" or "reasonable probability null hypothesis is true," perhaps more than students in a nonSBI course would.
- In Study 4, students struggled to answer why re-randomization is used in the *simulation*, tending to answer with why the (new) idea of random assignment is used in study design. This may be evidence of difficulty in distinguishing between the research study and the simulation model.

Table 2. Results from Study 3 (terminology), Spring 2022

Quiz question	Instructor 1 (in person)			Instructor 2 (online)		
	Definitions (n = 18)	Module (n = 26)	Video (n = 18)	Definitions (n = 7)	Module (n = 15)	Video (n = 18)
Variable	0.56	0.69	0.56	0.71	0.20	0.56
Sample	1.00	1.00	0.94	1.00	1.00	1.00
Population	0.89	0.92	0.89	1.00	1.00	1.00
Parameter	0.83	0.85	0.72	0.86	0.80	1.00
Statistic	0.94	0.85	0.78	0.86	0.93	1.00
Avg score (0-5)	3.67	3.62	3.33	4.43	3.93	4.56
	(0.69)	(0.80)	(1.09)	(0.79)	(0.80)	(0.51)

First 5 rows show proportion correct; last row is average score (and SD) on the 5-question quiz

Table 3. Results from Study 4 (comparing two proportions), Spring 2022

Quiz question	Instructor 1 (in person)			Instructor 2 (online)		
	Video1 (n = 16)	Video2 (n = 16)	zyBooks (n = 23)	Video1 (n = 14)	Video2 (n = 13)	zyBooks (n = 10)
Set up simulation (0-3)	2.78	3.00	2.90	2.91	2.82	3.00
Center of null	0.44	0.25	0.48	0.50	0.69	0.70
Why re-randomize	0.31	0.50	0.57	0.21	0.23	0.10
Each dot	0.81	0.81	0.70	0.64	0.77	0.60
Variable	0.31	0.25	0.44	0.36	0.62	0.30
Negative value	0.88	0.56	0.70	0.43	0.85	0.70
Two-sided *p*-value	0.25	0.00	0.26	0.14	0.31	0.50
Conclusion	0.38	0.38	0.30	0.43	0.54	0.30
Avg score (0-10)	6.15	5.75	6.33	5.63	6.82	6.20
	(2.03)	(1.44)	(2.02)	(1.83)	(1.78)	(2.10)
How helpful (1-5)	3.00	3.56	2.87	2.86	3.39	3.40
	(0.82)	(0.89)	(0.92)	(0.95)	(0.51)	(0.84)

First row is average score on 3-point matching question; next 7 rows are proportion correct, then average score (and SD) on a 10-point quiz, and average rating (and SD) on 1-5 Likert scale.

DISCUSSION

These studies have explored the feasibility of small classroom experiments exploring student learning from different initial presentations of key statistical ideas, especially as related to the

instruction of simulation-based inference. What did we learn from these studies and how can such implementations be improved in advance of more large-scale implementations?

Perhaps the richest source of data from these studies was a follow-up lab assignment given by one instructor in Study 2. A few weeks later in the course, the instructor revealed to students what the different treatments were and asked students to reflect on their own experiences, to conjecture which treatment would be more effective, and to analyze their results. Students were also asked about their fidelity to the treatments, which was incorporated into our data analyses by removing students who admitted to not following the instructions (e.g., finding a set of index cards at home to mark and shuffle), as well as to focus on the ethical nature of the study. Students then applied a two-sample *t*-test for two of the treatments and wrote a summary of their conclusions following the usual template in the course. This assignment fit perfectly into the regular lab assignments, with the added benefit of students talking from their own experiences and having additional ownership of the data. Student reflections of how to improve the study were generally of high quality.

Overall, we found that different instructors were able to integrate these common activities despite different course structures and modalities. The individual quizzes or pre-labs were consistent with other course assignments, although we do recognize that, especially when asked to listen to different instructors, the approach was not an exact match to the other topics. However, we feel that using one or two such activities was not a severe distraction in the course. Students did not express displeasure at being "experimented on" especially when given course credit and the opportunity to analyze the data. Several seemed to appreciate the opportunity to reflect on their own learning style as well. In the implementation of a fifth study, student pairs were literally side-by-side in a computer lab and exchanging notes on what they were learning from the different simulation strategies. In this study, it was also helpful for students to be trying the activities at exactly the same time, and assessments could be collected at the end of class period before outside resources could be consulted.

The main improvement we would suggest is collecting additional information on student fidelity. For example, although Canvas could record how long students spend on the quiz as a whole, these were not timed assessments, and students often left the assignment open for many hours while they reviewed course material. However, we did find (and removed) some students who had the quiz open for less time than the length of the video they were assigned. The data from zyBooks also indicated that many students did not fully participate in the "ask" portion of the module before moving to the actual quiz questions. As in the follow-up lab assignment, students could be asked, without penalty, how strictly they followed the instructions and sub-analyses could be performed. Some implementations were more transparent and asked students to "do their best" before receiving full credit. Further exploration is needed to find the optimal level of authenticity without encouraging duplicity or other large changes in how different students interact with the material. Restricting the interaction to class time is one option that would ensure more uniformity, but it also comes with other costs and can make assignment at the individual level more difficult.

Another issue to consider in designing such studies is how "bad" some treatments should be. We focused on genuine implementation questions we had (e.g., hands on simulation before computer simulation, students being one dot before proceeding to applet) but if the treatments are too similar, it is more difficult to find significant differences, even within a one-day stand-alone assignment. Still, we felt we could learn about small improvements (e.g., changing to people icons) that would likely not appear in course-level assessments. As much as possible, these multiple-choice questions, including how useful the student rated the activity, can be accompanied by open-ended responses as well.

Our final caution in developing such technology-based implementations is to be prepared for technology difficulty. For example, will students have difficulty following the links to videos housed outside of the course management system? What is the backup plan if the course management system is unavailable for a noticeable chunk of time? How do we track student use of outside resources without disrupting their normal practice? What if students submit accidently or fail to submit their assignment? What if the random assignment is not very well-balanced across the demonstrations?

CONCLUSIONS

In our experience, these classroom-based mini-experiments were moderately challenging to set up the first time but now that they have been created would be more straightforward to incorporate into future courses to better explore patterns, although we do find ourselves often wanting to tweak the

treatments each time. In fact, once different demonstrations are developed, they could all be presented to the students for them to explore their own learning styles. For example, giving students back-up reading(s) in case they do not feel comfortable after viewing the video alone. New technologies also provide more ability to track student progress (are they actually watching the videos) and to allow students to ask questions/provide feedback during the demonstrations to know where improvements are needed. Embedding concept-check questions within a video can also help students be more interactive in their learning of the material, but only if they are motivated to do so. We feel the abilities to change the demonstration and to quickly change the applet are critical for exploring such improvements, requiring the cooperation of the classroom instructors, curriculum developer, and applet programmer. Of course, this collaboration ideally occurs pre-course to allow for seamless integration of the demonstrations into the existing course structure. We did find the information gained from these mini-experiments on how and why to change students' instruction to key ideas to be fruitful, even in lieu of focus-groups or talk-aloud protocols.

As for simulation-based inference, in our experience students are often able to understand the notion of "could this have happened by chance alone?" However, getting students to explain in their own words what is meant by "this," "chance," and "alone," takes more repetition. These assignments have shown some evidence that how the material is introduced to students may help that initial development. In particular, using more concrete visuals and galvanizing more student engagement in developing the simulation models and even in the design of these studies may be key.

AKNOLWEDGEMENTS

We thank Julia Schedler and zyBooks for making it feasible for us to utilize specific subsections of the zyBooks text and Maddie Schroth-Glanz, Anelise Sabbag, Todd Swanson, and Jill VanderStoep and their students for participating in these efforts.

REFERENCES

Chance, B. L., & Garfield, J. B. (2001, Aug. 22–29). *New approaches to gathering data on student learning for research in statistics education* [Paper presentation]. 53rd Session of the International Statistical Institute, Seoul, Korea. www.isi-web.org/isi.cbs.nl/iamamember/CD2/pdf/751.PDF

Chance, B., & Rossman, A. (2006). Using simulation to teach and learn statistics. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education. Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS-7).* ISI/IASE. https://iase-web.org/documents/papers/icots7/7E1_CHAN.pdf?1402524965

Chance, B., Roy, S., McGaughey, K., Tintle, N., Swanson, T., & VanderStoep, J. (2021, August 11). *Using randomized experiments and common midterm questions to elucidate student learning trajectories from simulation-based inference curricula.* [Conference presentation]. Statistics, data, and the stories they tell. 2021 Joint Statistics Meetings Virtual Conference.

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., & Leader, S. (2022). Student performance in curricula centered on simulation-based inference: Final report. Manuscript in preparation.

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum. *Technology Innovations in Statistics Education*, *1*(1). https://doi.org/10.5070/T511000028

Scheaffer, R., Aliaga, M., Diener-West, M., Garfield, J., Higgins, T., Hilton, S., Hughes, G., Junker, B., Kepner, H., Kilpatrick, J., Lehrer, R., Lester, F. K., Olkin, I., Pearl, D., Schoenfeld, A., Shaffer, J., Silver, E., Smith, W., Speed, F. M., & Thompson, P. (2007). *Using statistics effectively in mathematics education research*. American Statistical Association.

Tintle, N., Chance, B. L., Cobb, G. W., Rossman, A. J., Roy, S., Swanson, T., & VanderStoep, J. (2020). *Introduction to statistical investigations* (2nd ed.). Wiley.

Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T., & VanderStoep, J. (2018). Assessing the association between pre-course metrics of student preparation and student performance in introductory statistics: Results from early data on simulation-based inference vs. nonsimulation based inference, *Journal of Statistics Education*. *26*(2), 103–109. https://doi.org/10.1080/10691898.2018.1473061

Tintle, N., VanderStoep, J., Holmes, V.-L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum, *Journal of Statistics Education*, *19*(1). https://doi.org/10.1080/10691898.2011.11889599