ASSESSING SIMULATION-BASED INFERENCE IN SECONDARY SCHOOLS

Soma Roy and Tessa Mcdonnel
Department of Statistics, Cal Poly – San Luis Obispo, CA 93401
soroy@calpoly.edu

Simulation-based inference methods continue to gain in popularity across grade levels and institutions. Despite their popularity, there continues to be a dearth of assessment data available demonstrating their efficacy across diverse student populations. We have recently collected comprehensive data about student performance in secondary school statistics courses using a variety of different curricular and pedagogical approaches, including, but not limited to, simulation-based inference. In this paper, we will summarize and discuss pre-course and post-course conceptual understanding in aggregate and on topic specific scales between high school courses using simulation-based methods and those that are not.

INTRODUCTION

The demands for a statistically literate society are increasing (McKinsey report 2011), and the introductory statistics course ("Stat 101") remains the primary venue for learning statistics for the majority of high school and undergraduate students. The last few decades have seen much fruitful activity in the areas of pedagogy and assessment, with the focus slightly shifting, in the recent years, towards rethinking the content of this course, and more specifically on using simulation-based methods, including bootstrapping and permutation tests, to illustrate core concepts of statistical inference within the context of the overall statistical investigative process. This new focus presents an opportunity to address documented shortcomings in the standard Stat 101 course, such as helping students to see the big picture early and throughout the course, and improving students' statistical thinking over mere knowledge of procedures.

Our group has developed and implemented one of the first comprehensive curricula (Tintle et al. 2016) that (a) emphasizes the core logic of inference using simulation-based methods with an intuitive, cyclical, active-learning pedagogy, and (b) emphasizes the overall process of statistical investigations, from asking meaningful research questions and collecting data, through making inferences and drawing conclusions. Improved conceptual understanding (Tintle et al. 2011) and retention of inference and study design concepts (Tintle et al. 2012), which had been observed with early versions of the curriculum at a single institution are now being evaluated at dozens of institutions across the country with thousands of students. Encouraging preliminary results continue to be observed. These materials are now available as the textbook *Introduction to Statistical Investigations* (*ISI*), published by John Wiley and Sons (2016).

This paper will focus on our findings from implementing simulation-based curricula at the high school level. In the next section, we briefly describe our curriculum, our assessment instrument, our sample, and the statistical analyses for the data collected. Results are provided in the following section, with a discussion of follow-up questions in the last section.

METHODS

Our Curriculum and Materials

In our textbook (*ISI*, 2016), we introduce statistical inference in week 1, using simulation and randomization methods to do so. We use a spiral approach to present the statistical investigation process that begins by asking the research question, moves on to designing the study and collecting data, analyzing the data, and ends with formulating conclusions and suggesting follow-up research questions. The philosophy behind this is to expose students to the logic of statistical inference early, using simulation-based methods, and give them time to develop and strengthen their understanding as they repeatedly revisit the logic of inference, specifically the concepts of *p*-value and confidence interval, in new scenarios. This approach makes use of modern computing power and puts the logic of statistical inference at the center of the curriculum, as advocated by Cobb (2007). Our materials also have the additional key features of integrating exposition, examples, and active explorations, integrating easy-to-use web-based applets (Rossman/Chance applets, www.rossmanchance.com/applets) for carrying out analyses, and always using real data from genuine studies that matter.

Assessment Instrument

To assess the effectiveness of the simulation-based curriculum at the high school level we used a tool that has been recently developed by our group. This tool uses a few items from the Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS; delMas et al. 2007); a few items were modified to clarify what it was assessing, and a few questions were omitted because students consistently perform very well on these items on pre-tests. A few additional items were developed to target areas that were not addressed thoroughly by the existing items. We developed this assessment tool to be one that any introductory statistics student should be able to do well on, regardless of the type of course (simulation-based or traditional) that they were taking. More information about this tool as well as results on its validity and reliability across over 3000 students at several institutions, and information about external factors associated with student performance (e.g., test setting and question order) are discussed in Tintle et al. (2018).

This assessment tool (with 24 questions) was combined with the Student Attitudes Towards Statistics (SATS-36) developed by Schau (2003) into one instrument, and the entire instrument was administered once towards the beginning of the course, and once towards the end. This enabled us to measure gain in a student's statistical knowledge, as well as any change in their attitudes towards statistics over the course. The results for the attitudes survey will be discussed elsewhere.

Additionally, each participating instructor was also asked to respond to a survey about themselves and various aspects of their teaching, such as, number of years of teaching experience, whether or not use simulation-based curriculum, and whether class is a full-year class or a semester class. Similarly, students were asked to report among other things, their sex, and current grade point average (GPA).

The Sample

We recruited a mix of instructors so that we had both users of simulation-based curricula, as well as those who used more traditional (non-simulation-based) curricular materials. Because all the instructors in our sample who were using simulation-based curricula at the secondary level turned out to be using the ISI materials, we will refer to these instructors and their students as the ISI group, and the other group as the NonISI. The students of these instructors are our observational units. Overall, we collected data on students from 28 different high schools from across the United States (16 states), in the academic year 2016-17. Among the 28 schools, 19 are public schools, six are private schools, one is a public charter, one is a private charter, and one is a private boarding school. We ended up having usable data (where students had taken both the pre-test and post-test, as discussed in the next section) from 24 instructors, for just over 630 students.

Table 1 – Descriptive statistics on the instructors

	ISI	**NonISI**	**Overall**
N	10	14	24
Female instructor (%)	50%	50%	41.7%
Mean number of years of teaching intro statistics (SD)	10.1 (6.2)	11.1 (4.8)	10.7 (5.3)
Whether class was full year	50.6%	95.2%	91.1%

Table 1 shows overall summary statistics for the 24 instructors, as well as summary statistics separated by whether the instructor was using simulation-based curriculum materials (more specifically ISI), or not (NonISI). Both groups of instructors had, on average, a similar number of years of experience teaching introductory statistics, and half in each group were females. Note that only about half the instructors in the ISI group taught the statistics class as a full year class, compared to about 95% of the

NonISI instructors. Many of the full year classes were an Advanced Placement course in Statistics, which is typically a year-long curriculum, developed by the College Board, and students who pass a national exam at the end of the year receive university credit for the course.

The summary statistics on the students of the 24 participating instructors are provided in Table 2. Only students who took both the pre-test and the post-test, and opted to let us use their data for research purposes, are included in the analysis dataset. It is notable that the reported incoming GPA was about the same for both groups, on average.

Table 2 – Descriptive statistics on the students

	ISI	**NonISI**	**Overall**
N	196	435	631
Female student (%)	50.9%	58.2%	53.2%
Mean Incoming GPA (n = *554*), and (SD)	3.78 (0.24)	3.68 (0.31)	3.71 (0.29)

Statistical Analyses

We chose to use gain as well as achievable gain (which is defined as the ratio of actual gain in score from before to after, to the maximum possible gain from before to after; Hake, 1998, 2002), to measure change in students' statistical knowledge over the course. To investigate whether there were any effects of curriculum, we ran a hierarchical model to predict students' gain as well as achievable gain, and accounted for whether or not simulation-based materials (more specifically ISI) were used, as well as instructor effects, pre-test score, sex of student, incoming GPA, and whether and what type of incentive was used to encourage students to take the pre- and post-tests.

ASSESSMENT RESULTS

The overall mean gain (post-test score minus pre-test score) was calculated to be 0.150 (SD = 0.157), which implies that from before to after, on average, students were getting the answers correct to roughly 3.8 additional questions on the statistical knowledge assessment. The overall mean achievable gain in our sample was found to be 0.273, with a standard deviation of 0.299.

The parameter estimates from our hierarchical linear model are shown in Table 3. We found that, on average, students' gain was 0.024 points higher for those who use the ISI (simulation-based) curriculum compared to non-simulation-based curricula; however, this was not a significant difference after adjusting for other predictors. Males had significantly higher gains (p-value < 0.001), on average, compared to females. Incoming GPA was found to have a significant positive association with gain (p-value <0.001), whereas pre-test score was found to have a significant negative association with gain (p-value < 0.001). Also, note that students offered a large incentive (e.g., chance to enter a raffle for a $50 Amazon.com gift card) showed a significantly (p-value < 0.05) higher average gain, compared to when no incentive was offered. We also found that compared to Fall semester classes, full year (p-value < 0.05) and Spring (p-value < 0.10) semester classes had higher gains, on average.

The results from using achievable gain as the response variable were similar to that of using gain, except that the difference between the large incentive and no incentive group was more statistically significant (p-value < 0.01) than with gain as the response variable.

The 24 questions on the statistical knowledge assessment tool were divided into five subscales: data collection and scope of conclusions, descriptive statistics, confidence interval, significance tests, and simulation and sampling distributions. As a follow-up, we took a closer look at how students did on these subscales, and compared gain across curricula. Table 4 shows the adjusted mean differences in percentage correct (pre to post) comparing students who used the ISI (simulation-based) curricular materials to those who used other (non-simulation-based) materials. We found that students in the ISI (simulation-based) curriculum compared to non-ISI had higher average adjusted differences in percentage correct on the data collection, descriptive statistics, simulation/sampling variability, and significance tests subscales, with only the results from the significance tests subscale being significant. On the confidence interval subscale students in the non-ISI curricula had a higher, though not statistically significant, average adjusted difference in percentage correct.

Table 3 – Parameter estimates from a hierarchical linear model where models 1 and 2 use gain and achievable gain as response variables, respectively

	Model 1 Estimates		**Model 2 Estimates**	
Intercept	-0.190	*	-0.660	**
Whether instructor is a male (ref = female)	-0.015		-0.027	
Duration/time of course (ref = Fall semester)				
Full year class	0.098	*	0.178	*
Spring semester	0.081	.	0.166	.
Type of incentive (ref = none offered)				
Large incentive offered	0.106	*	0.215	**
Some incentive offered	0.020		0.054	
Other (missing, etc.)	0.019		0.073	
Whether student is a male (ref = female)	0.0495	***	0.096	***
Incoming GPA	0.111	***	0.209	***
Pre-test score	-0.622	***	-0.574	***
ISI curriculum (ref = non-ISI curriculum)	0.024		0.032	

Note: *** indicates p-value < 0.001; ** indicates p-value < 0.01; * indicates p-value < 0.05; . indicates p-value < 0.10.

Table 4 – Adjusted mean gain (± SD) separated by curriculum

	ISI	**NonISI**
n	196	435
Achievable gain	0.214 (0.051)	0.182 (0.058)
Total gain (out of 24 questions)	0.124 (0.025)	0.098 (0.025)
Score on subscale		
Data collection/Scope of conclusions (4 questions)	0.104 (0.041)	0.074 (0.048)
Descriptive statistics (5 questions)	0.077 (0.049)	0.080 (0.056)
Confidence intervals (5 questions)	0.152 (0.036)	0.199 (0.044)
Significance Tests (7 questions)	0.146 (0.037)	0.063 (0.043)*
Simulation/Sampling variability (3 questions)	0.143 (0.051)	0.098 (0.060)

Note: on the subscales, statistics are average adjusted differences in percentage correct (SD); * indicates that the difference is significant at 5% level of significance.

A similar analysis of assessment data on post-secondary students indicates a somewhat stronger positive significant effect of simulation-based curricular materials on gain (Chance et al. 2018).

DISCUSSION AND CONCLUSIONS

Our results suggest that, with regard to performance on a statistical assessment, among secondary (that is, high school) students there is a slight (though not statistically significant) positive effect of the ISI (simulation-based) curriculum compared to “traditional” curriculum that uses theory-based inferential procedures exclusively. This is in contrast to the somewhat significant positive effect that was seen among tertiary students (Chance et al. 2018). On the topic subscales, high school students using the simulation-based curriculum had significantly higher gains on questions on statistical significance, compared to those not using simulation-based curricula.

Our next steps involve, first, investigating whether the small, though not significant, effect of curriculum holds up when we combine data from other institutes and across different years in a more robust study. We would also like to understand better the difference in student performances on the various subscales, across the different curricula, and across secondary versus tertiary levels of education. We will also be considering adjusting our models for previous statistics knowledge, and interactions of type of curriculum with student characteristics.

ACKNOWLEDGEMENTS

We thank the National Science Foundation for two grants (1140629 and 1323210) that have supported the development and evaluation of this curriculum.

REFERENCES

Advanced Placement Statistics Course, https://apstudent.collegeboard.org/apcourse/ap-statistics/course-details (last accessed March 9, 2018)

Chance, B., Mendoza, S., and Tintle, N. (2018). Student Gains in Conceptual Understanding in Introductory Statistics with and without a Curriculum Focused on Simulation-based Inference, *Proceedings of the 10th International Conference on Teaching Statistics.*

Cobb, George W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum?, *Technology innovations in statistics education 1*(1): 1-15.

delMas, R., Garfield, J., Ooms, A., and Chance, B. (2007). Assessing Students' Conceptual Understanding After a First Course in Statistics, *Statistics Education Research Journal*, *6*, 28–58. Available at http:// www.stat.auckland.ac.nz/»iase/serj/SERJ6%282%29_delMas.pdf

Hake, R. R. (1998). Interactive Engagement Versus Traditional Methods: A Six-Thousand Student Survey of Mechanics Test Data for Introductory Physics Courses, *American Journal of Physics*, *66*, 64–74. Available at http://www.physics.indiana.edu/#sdi

Hake, R. R. (2002). Normalized Learning Gain: A Key Measure of Student Learning (Addendum to Melzer, D. E. (2002), The Relationship Between Mathematics Preparation and Conceptual Learning Gains in Physics: A Possible 'Hidden Variable' in Diagnostic Pretest Scores), *American Journal of Physics*, *70*, 1259–1267

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh,C., and Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute. Available at https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Tintle, N., (2018). Development of a tool to assess students' conceptual understanding in introductory statistics, *Proceedings of the 10th International Conference on Teaching Statistics.*

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., and VanderStoep, J. (2016). *Introduction to Statistical Investigations*, 1st edition. Hoboken, New Jersey: John Wiley and Sons.

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V., and Swanson, T. (2012). Retention of Statistical Concepts in a Preliminary Randomization-Based Introductory Statistics Curriculum. *Statistics Education Research Journal* 11(1).

Tintle, N., VanderStoep, J., Holmes, V., Quisenberry, B., and Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, *19*(1).