STUDENT GAINS IN CONCEPTUAL UNDERSTANDING IN INTRODUCTORY STATISTICS WITH AND WITHOUT A CURRICULUM FOCUSED ON SIMULATION-BASED INFERENCE

Beth Chance[1], Stephanie Mendoza[1], and Nathan Tintle[2]
[1]Department of Statistics, Cal Poly – San Luis Obispo, CA 93401
[2]Department of Math, Computer Science & Statistics, Dordt College, Sioux Center, IA
bchance@calpoly.edu

Using "simulation-based inference" (SBI) such as randomization tests as the primary vehicle for introducing students to the logic and scope of statistical inference has been advocated with the potential of improving student understanding of statistical inference, as well as the statistical investigative process as a whole. Moving beyond the individual class activity, entirely revised introductory statistics curricula centering on these ideas have been developed and tested. In this presentation we will discuss three years of cross-institutional tertiary-level data in the United States comparing SBI-focused curricula and non-SBI curricula (roughly 15,000 students). We examine several pre/post measures of conceptual understanding in the introductory algebra-based course, using hierarchical modelling to incorporate student-level, instructor-level, and institutional-level covariates.

INTRODUCTION

The demands for a statistically literate society are increasing, and the introductory statistics course ("Stat 101") remains the primary venue for learning statistics for the majority of secondary and tertiary students. Despite three decades of very fruitful activity in the areas of pedagogy, assessment results have not shown dramatic gains in student understanding or appreciation of statistics. Surveys have also shown that although many instructors have made changes in their courses with respect to technology, use of genuine data, and projects, the changes have been slow in developing (Garfield, 2000).

More recent calls for reform have focused on not only pedagogy and assessment methods, but also course content. One example is Cobb's 2007 call to center the introductory curriculum around the reasoning and logic of statistical inference, rather than the normal distribution. Recent technology tools enable students to use simulation-methods, e.g., bootstrapping and randomization tests, to develop confidence intervals and p-values with minimal mathematical distractions. This allows students to focus more closely on the statistical investigation process as a whole rather than seeing data collection, data exploration, probability, and statistical inference as unrelated units of instruction. Preliminary assessment results, primarily at single institutions, have shown promising benefits to this approach (e.g., Tintle et al., 2011, 2012; Beckman, delMas, & Garfield, to appear; Chance & McGaughey, 2014; Hildreth, Robison-Cox, & Schmidt, to appear 2018; Maurer & Lock, 2015).

The last few years have seen several full curricula/textbooks for introductory algebra-based statistics developed which focus on "simulation-based inference" (SBI) including *Introduction to Statistics Investigations* (ISI; Tintle et al.), *Statistics: Unlocking the Power of Data* (Lock5; Lock et al.), *Statistical Reasoning in Sports* (Tabor & Franklin), *Statistical Thinking: A simulation approach to modeling uncertainty* (Zieffler & Catalysts for Change), and *Introductory Statistics with Randomization and Simulation* (OpenIntro). These textbooks vary in the depth of their treatment of simulation-based methods, the choice of simulation methods, the sequencing of topics, and the technology tools, but all provide an alternative introduction to statistical reasoning that focuses on using simulation to help students understand randomness and inference.

As part of an NSF grant for developing the ISI curriculum, our team also sought to assess students' development of conceptual understanding in the introductory course, comparing the SBI courses to other curricula across scores of institutions. After pilot-testing the instrument, we now have three years of data across approximately 15,000 students. The discussion below will describe our methods and implementation of our assessment project. These data allow us to compare student pre/post performance and attitudes across several curricula, while incorporating student level, section level, and instructor level covariates. Below we focus on these textbook comparisons. More

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2018 ISI/IASE]

details on the hierarchical regression models, which can help identify which factors are most highly related to improved performance and perhaps which teaching decisions have the most impact, will be provided in the presentation.

SBI CURRICULA

The main distinction of SBI curricula is using simulation as the primary vehicle for introducing ideas in statistical inference. In the "Lock5" curricula, students learn methods of bootstrapping and randomization tests for confidence intervals and tests of significance before carrying out any formal inference procedures. Free on-line applets (StatKey) are used to generate bootstrap and null distributions to estimate percentile intervals and p-values. This understanding is then built upon as students move through more standard "theory-based" inference procedures. Similarly, the ISI curricula uses free on-line applets (www.rossmanchance.com/applets/) focusing on estimating p-values and standard errors. One distinction from the Lock5 text is this SBI material is discussed very early in the course (e.g., week 1) as one piece of the overall statistical investigation process (Roy, 2014), and a spiral approach is used to revisit these ideas with new scenarios (e.g., comparing groups, association). The CATALST curriculum (Garfield, delMas, & Zieffler, 2012) uses Tinkerplots^{TM} to explore the same ideas, focusing on chance models and simulation, however the curriculum does not cover as many of the traditional testing procedures. Other instructors have developed their own materials that are hybrids of these approaches (e.g., Hildreth, Robison-Cox, & Schmidt, to appear). In the analyses below, we differentiate between ISI, ISI-first (instructors using the ISI materials for their first time), other SBI (e.g., Lock 5 and CATALST), non-SBI, and "other" (materials developed and used at the individual institution) courses. (We did not review the other materials in detail to classify them, but several are known to be heavily influenced by SBI.) The non-SBI courses were also divided into what the first author considered "GAISE-compliant" and "non GAISE-compliant" (non SBI-2) textbooks (http://www.amstat.org/asa/files/pdfs/ GAISE/GaiseCollege_Full.pdf).

METHODS

A 32-question multiple-choice concept inventory was developed by the ISI author team, modelled after the CAOS instrument from the University of Minnesota (delMas, Garfield, Ooms, and Chance, 2007). Topics on this inventory include data collection, simulation/probability, descriptive statistics, confidence intervals, significance tests, and scope of conclusions. This inventory was combined with the SATS-36 instrument (Student Attitudes Toward Statistics; Schau, 2003) into one instrument. (See Tintle et al., 2018 for a report on the validity and reliability of the instrument.) Faculty were recruited to participate using email listservs (e.g., ASA Section on Statistics Education, Isolated Statisticians, SIGMAA on statistics education). Faculty were asked to give their students the combined instrument through SurveyMonkey during the first and last weeks of the term. Students were given the option to opt out of their results being used for research purposes. Faculty were encouraged to give their students some incentive (e.g., credit on a homework assignment) for participating. Faculty were also asked to complete a survey about their own background and teaching methods as well as details of the course (e.g., number of students, meeting time, use of active learning, familiarity with the GAISE guidelines, type of institution).

IMPLEMENTATION

The primary response of interest is students' performance on the concept inventory. To adjust for pre-test scores and possible ceiling effects, we utilize *achievable gain* = *gain*/(1–*pre*) as our measure of student improvement (aka "single-student normalized gain", e.g., Colt, Davoudi, Murgu, & Zamanian Rohani, 2011; Hake, 1998). Students with an achievable gain of -2 and lower (e.g., 78% correct on the pre-test, 33% correct on the post-test) were removed from the analyses. Students who opted out or who answered less than 60% of the questions on either test were also removed. (This includes students who took a "condensed" version of the post-test with 10 questions.) We felt these observations were not trustworthy measures of student knowledge.

Examples of the lengthy data cleaning tasks include reconciling discrepancies in demographic data between pre and post administrations (e.g., students changing sex, self-reported GPAs, age) and tracking students who changed instructors after the pre-test. Students also self-reported SAT or ACT scores which were converted into a *z*-score based on the means and standard deviations of each scale in our dataset (which were similar to nationally reported values). Text responses to numeric questions were converted (e.g., "I think my GPA is around 3.2"). The cleaned data set was merged with a cleaned version of the teacher inventory.

The instrument was field tested in Fall 2012 and used with approximately 2000 students in 2013/2014 (after cleaning), with mostly ISI instructors. In Year 1 of the study (2014/2015), after a few modifications, the instrument was given to over 3,000 students across 70 instructors and 38 institutions. In Year 2, this was expanded to 95 instructors at 57 institutions. In Year 3, this was expanded to over 8,000 students for 95 instructors across 78 institutions.

After data cleaning, the final pre/post data set for Year 3 consisted of 4440 students, 227 instructor-terms (some instructors participated in both fall and spring), and 70 institutions, mostly universities (*n* = 2305) and four-year colleges (*n* = 1182) with some community or two-year colleges (*n* = 263) and high schools (*n* = 606, year-long course). These gave us 104 sections using *ISI* (49 for the first time), 45 using other SBI texts, and 106 not using an SBI curriculum (39 classified as not GAISE compliant). One section using a text with a calculus prerequisite was removed.

RESULTS

Gains in conceptual understanding

Table 1 shows the mean achievable gain scores across the textbooks for the three years. It's important to note that the achievable gains are modest, but we see a consistent pattern across the years: a tendency for higher gains with the ISI textbook, regardless of whether or not the instructor is using the text for the first time, and a similar but slightly lower improvement with the other SBI texts. Typically the means are lower for the non-SBI students, especially those in courses using textbooks that were not considered GAISE compliant. The last row shows that in Year 3 scores generally declined if the high school (year-long) sections are removed from the analysis, but also showing stronger textbook effects. There is still considerable within-section variability (Figure 1) that we explored further using hierarchical regression models incorporating student and instructor level data. This pattern has held true even after adjusting for student background (e.g., SAT/ACT *z*-score, GPA, pre-test score).

Table 1. Mean achievable gain scores for 2014/2015 – 2016/2017 school years, across textbooks

	Overall	ISI	ISI 1st	Other SBI	Not SBI	Not SBI-2
Year 1	0.170	0.216	0.214	0.195	0.124	0.061
Year 2	0.108	0.170	0.160	0.132	0.076	0.030
Year 3	0.169	0.191	0.191	0.169	0.166	0.101
Year 3 (no HS)	0.147	0.179	0.177	0.169	0.104	0.060

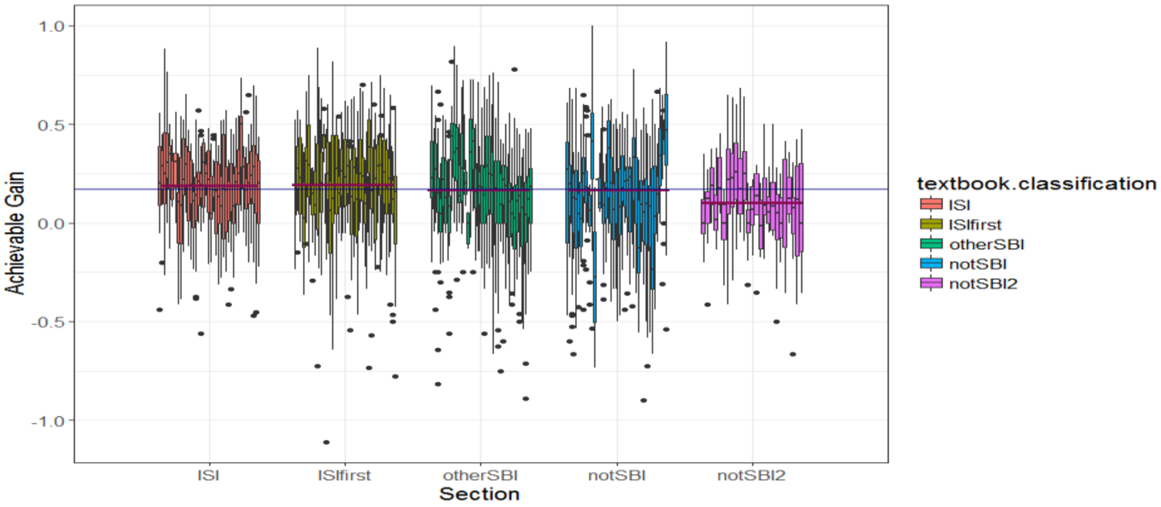

Figure 1. Boxplots of achievable gain by textbook for 227 instructor-terms (Year 3).
Red lines indicate averages by textbook classification.

One observation of interest from the hierarchical models is that Carnegie classification (e.g., primarily undergraduate institution vs. research university) explains more variation in the data than the section or instructor or institution. For student level variables, the strongest predictors of achievable gain include pre-test, SAT/ACT *z*-score, and GPA. An analysis by Tintle et al. (to appear) conditioning on "student preparation" (standardized pre-college test scores) showed equal benefits to both more quantitatively and less quantitatively mature students. A preliminary analysis (Chance, Wong, & Tintle, 2016) showed evidence of a quadratic relationship with GPA as well as interactions with student and instructor sex. These relationships will be explored further in subsequent analyses.

Student attitudes

The SATS has been used by several authors to explore changes in student attitudes in introductory statistics course. The SATS consists of several questions within 6 subscales: affect, cognitive competence, difficulty, effort, interest, and value. However, few studies, with a few notable exceptions, have found large improvements or even changes in student attitudes from the first course. (See for example the 2012 *SERJ* special issue, 11(2).) Table 2 shows the gain (post-pre) in the attitude subscales across the textbooks for Year 3. (Cronbach α values were quite comparable to other studies as well.)

Table 2. Year 3 attitude changes from SATS across textbook type (post – pre), *n* = 4414

	ISI	ISI 1st	Other SBI	Not SBI	Not SBI-2	Own
n	600	662	1103	1279	291	479
Affect	0.235	0.074	0.174	0.033	0.021	0.080
Cog Comp	0.174	-0.043	-0.018	-0.077	-0.194	0.048
Difficulty	0.405	0.227	0.236	0.189	0.190	0.314
Effort	-0.967	-1.01	-0.920	-1.03	-1.23	-0.943
Interest	-0.565	-0.601	-0.449	-0.515	-0.740	-0.728
Value	-0.250	-0.312	-0.230	-0.270	-0.372	-0.371

The SBI courses do show more positive gains on the affect and cognitive competence scales. The positive changes in difficulty imply the students tended to find the course less difficult than expected (more so for the SBI courses). The negative changes in effort imply the students ended up putting in less work than they anticipated. Unfortunately, we also see negative changes in the interest and value scales, though less so with the SBI curricula.

Question by question comparisons on concept inventory

Below (Table 3) we highlight some individual questions from our concept inventory using the 2016-2017 data (4646 students on the pre-test, 5166 students on the post-test). One concern with a course focused on simulation-based inference is a loss of time on other topics (e.g., descriptive statistics). Our results indicate that most recently, students in simulation-based inference courses show similar gains in understanding as students in the non-simulation-based curricula. For example, Question 22 on the concept inventory requires students to realize that they can compare two distributions of a quantitative variable even though the groups have unequal sample sizes. Students entering the college course show proficiency on this topic, increasing by the end of the course across all the curricula.

As hoped, students in the SBI curricula do show large gains in questions related the reasoning of statistical inference. Question 27 on the concept inventory asks students whether a research is hoping for a large or small p-value to establish their hypothesis. Students are near 50/50 entering the course, but students leave the course with better understanding of the goals of statistical inference, especially in the simulation-based curricula. Students also demonstrated improvement in recognizing a correct p-value interpretation (Q29) as well as invalidating an incorrect interpretation (Q28, probability the null is true).

Some questions still showed little improvement or even decreases in performance. For example, Question 24 provides an insignificant p-value and asks whether that provides evidence for the null. All groups performed more poorly on this question on the posttest. Question 26 asked students to estimate the sample sized needed for a 3% margin of error in a survey of US adults (population size 310 million). Students improved but still performed very poorly on this question.

Students in the SBI courses showed more improvement in (informally) recognizing a result of 13 successes in 15 attempts as unlikely to happen by random chance (Q43) and in recognizing a correct description of a simulation (Q38) but did not outperform their peers in recognizing an inappropriate simulation plan (Q37, repeat the experiment many times).

Table 3. Question by question comparisons of interest (proportion correct responses)

		ISI	ISI 1st	Other SBI	Non SBI	Non SBI2
Q22: comparing distributions	Pre	0.69	0.71	0.70	0.73	0.68
	Post	0.86	0.85	0.85	0.81	0.69
Q27: want large or small p-value	Pre	0.42	0.36	0.48	0.40	0.29
	Post	0.92	0.88	0.83	0.77	0.60
Q29: correct p-value interpretation	Pre	0.43	0.42	0.42	0.44	0.35
	Post	0.65	0.65	0.59	0.52	0.44
Q28: p-value is prob of null	Pre	0.56	0.51	0.58	0.57	0.44
	Post	0.83	0.82	0.79	0.76	0.57
Q24: evidence for null	Pre	0.83	0.80	0.78	0.83	0.81
	Post	0.65	0.68	0.66	0.67	0.60
Q26: margin of error	Pre	0.13	0.12	0.16	0.14	0.13
	Post	0.24	0.27	0.24	0.33	0.21
Q43: 13/15	Pre	0.41	0.40	0.40	0.40	0.35
	Post	0.62	0.56	0.52	0.44	0.34
Q38: correct simulation	Pre	0.56	0.64	0.56	0.52	0.47
	Post	0.88	0.87	0.76	0.65	0.53
Q37: incorrect simulation	Pre	0.39	0.37	0.39	0.44	0.35
	Post	0.39	0.36	0.44	0.46	0.33

These results suggest that while students in SBI courses perform on par with their peers on many questions and above their peers on others, there are still areas for improvement across all the curricula. In particular, the roles of the null hypothesis and sample size in the simulation models needs to be better addressed.

DISCUSSION

Across three years of data (about 5,000 students per year after cleaning), we are seeing some consistent comparisons of the curricula, with courses that fully integrate simulation-based inference seeing slightly higher gains in student achievement on average. These differences have also been consistent in hierarchical models that adjust for student-level (e.g., GPA, ACT/SAT scores) and instructor-level data discussed elsewhere. These trends have also been consistent across the types of questions on the concept inventory. Still, there remains substantial within-section variation for further exploration. Further analyses will also consider

- Which student and instructor level variables explain the most variation in achievable gain? Are there interactions, e.g., between student pre-attitude and instructor gender or background?
- When the concept and attitude surveys are given in one instrument, does it matter which is given first? What is the role of incentives on student performance on these instruments?
- Do we see similar differences across textbooks four months after the course?
- Can we conjecture a learning trajectory of students understanding of statistical inference based on the number of exposures to simulation-based inference?

We also plan to make our dataset available to other researchers after the final cleaning steps.

ACKNOLWEDGEMENTS

This research was supported by NSF grants NSF/TUES/DUE-Phase II # 1323210. We would also like to thank Cal Poly undergraduates who were indispensable in the data cleaning phases of this research, including Brannden Moss, Danielle Simmons, and Rowan Kardos whose summer research was supported by the Bill and Linda Frost Fund, as well as the ISI team, including Cindy Nederhoff. For more discussion of issues related to implementing Simulation-Based Inference, please see the blog: www.causeweb.org/sbi.

REFERENCES

Beckman, M., delMas, R., & Garfield, J. (to appear). Cognitive Transfer Outcomes for a Simulation-Based Introductory Statistics Curriculum, *Statistics Ed Research Journal*, 16(2).

Chance, B., & McGaughey, K. (2014). Impact of a Simulation/randomization-Based Curriculum on Student Understanding of P-Values and Confidence Intervals. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics* (ICOTS9, July, 2014), Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

Chance, B., Wong, J., & Tintle, N. (2016). Student Performance in Curricula Centered on Simulation-Based Inference: A Preliminary Report. *Journal of Statistics Ed*, *24*(3), 114-126.

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

Colt, G. C., Davoudi, M., Murgu, S., & Zamanian Rohani, N. (2011), Measuring Learning Gain During a One-day Introductory Bronchoscopy Course. *Surgical Endoscopy, 25*(1), 207-216.

delMas, R., Garfield J., Ooms, A., & Chance, B. (2007), Assessing Students' Conceptual Understanding After a First Course in Statistics. *Statistics Ed Research Journal*, *6*(2), 28–58.

Garfield, J. (2000). An evaluation of the impact of statistics reform: Final Report. National Science Foundation (REC-9732404).

Garfield, J., delMas, R., and Zieffler, A. (2012). Developing Statistical Modelers and Thinkers in an Introductory, Tertiary-level Statistics Course. *ZDM – The International Journal on Mathematics Education*, *44*(7), 883–898.

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.

Hildreth, L., Robison-Cox, J., & Schmidt, J. (2018). Comparing Student Success and Understanding in Introductory Statistics Under Consensus and Simulation-Based Curricula, *Statistics Education Research Journal*, *17*(1).

Maurer, K., & Lock, E. (2015). Bootstrapping in the Introductory Statistics Curriculum, *Technology Innovations in Statistics Education*, *9*(1).

Roy, S., Rossman, A., Chance, B., Cobb., G., VanderStoep, J., Tintle, N., Swanson, T. (2014). Using Simulation/Randomization to Introduce p-values in Week 1. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics* (ICOTS9, July, 2014), Voorburg,

Schau, C. (2003). *Survey of Attitudes Toward Statistics* (http://evaluationandstatistics.com/)

Tintle, N., Topliff, K., VanderStoep, J., Homes, V-L., & Swanson, T. (2012). Retention of Statistical Concepts in a Preliminary Randomization-Based Introductory Statistics Curriculum. *Statistics Education Research Journal*, *11*(1), 21-40.

Tintle, N., & VanderStoep, J. (2018). Development of a Tool to Assess Students' Conceptual Understanding in Introductory Statistics, *Proceedings of ICOTS-10*, Kyoto, Japan.

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).

Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T., & VanderStoep, J. (to appear). 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. Non-Simulation Based Inference. *Journal of Statistics Education*.