STUDENT PERFORMANCE IN CURRICULA CENTERED ON SIMULATION-BASED INFERENCE

BETH CHANCE
California Polytechnic State University
bchance@calpoly.edu

NATHAN TINTLE
University of Illinois, Chicago
Nathan.Tintle@isi-stats.com

SHEA REYNOLDS
California Polytechnic State University
shereynolds@csumb.edu

AJAY PATEL
California Polytechnic State University
ajay12698@yahoo.com

KATHERINE CHAN
California Polytechnic State University
kchan78@calpoly.edu

SEAN LEADER
California Polytechnic State University
spleader@calpoly.edu

ABSTRACT

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 and the statistical investigative process. Moving beyond the individual class activity, entirely revised introductory statistics curricula centering on these ideas have been developed and tested. Preliminary assessment data have been mostly positive. In this paper, we discuss three years of cross-institutional tertiary-level data from the United States comparing SBI-focused curricula and non-SBI curricula (86 distinct institutions). We examined several pre/post measures of conceptual understanding in the introductory algebra-based course using multi-level modelling to incorporate student-level, instructor-level, and institutional-level covariates. We found that pre-course student characteristics (e.g., prior knowledge) were the strongest predictors of student learning, but also that textbook choice can still have a meaningful impact on student understanding of key statistical concepts. In particular, textbook choice was the strongest "modifiable" predictor of student outcomes of those examined, with simulation-based inference texts yielding the largest improvements in student learning outcomes. Further research is needed to elucidate the aspects of SBI curricula that contribute to observed student learning gains.

Keywords: *Statistics education research; Randomization tests; Multi-level models*

1. INTRODUCTION

The demands for a statistically literate society are increasing, and the introductory statistics course remains the primary venue for learning statistics for many secondary and tertiary students. As statistics and data science content reaches broader audiences of students, statistics educators have argued for changes in technology, pedagogy, and content (e.g., Cobb, 1992; Moore, 1997). Around this time,

Statistics Education Research Journal, 21(3), Article 4. https://doi.org/10.52041/serj.v21i3.6

Garfield (1995; see also Garfield & Ben-Zvi, 2007) illustrated how theories of learning could be applied to statistics education, namely students actively constructing their knowledge through active learning, practice, technology, and consistent and helpful feedback. These recommendations culminated in the first *GAISE (Guidelines for Assessment and Instruction in Statistics Education)* college report (Aliaga et al., 2005), but Garfield et al. (2002) found that although many instructors had made changes in their courses with respect to technology, use of genuine data and projects, the changes were slow in developing. Assessment results (delMas et al., 2007) often found students still struggling with key ideas, including how to interpret and correctly use *p*-values. Anecdotally, students reviewing for the final exam in the course would still ask instructors, "now that I have my *p*-value, do I want it to be large or small?" indicating they were learning the mechanics of the content but not understanding the big ideas of the course.

More recent calls for reform have focused on not only pedagogy and assessment methods, but also course content. Studies have shown that engaging students in modeling and generating distributions (e.g., Doerr & English, 2003; Konold et al., 2007) helps them understand randomness and chance, as well as providing explicit experience with sampling variability. As noted by Lee et al. (2015), prior research (e.g., Garfield et al., 2012; Lane-Getaz, 2007; Saldanha & Thompson, 2002) found that a three-tiered approach to constructing an inference problem, problems/models, repeated samples, and sampling distributions, appears to help students and teachers better conceptualize statistical inference. The argument for heavier use of simulation to teach the logic and scope of inference was brought to the forefront by Cobb's (2007) call to center the introductory curriculum around the reasoning and logic of statistical inference, rather than around the normal distribution. Cobb advocated persuasively for using repeated sampling and re-randomization, often starting with tactile simulations, to introduce students to the logic of statistical inference before proceeding to more traditional inference procedures. This supports students in developing conceptual understanding of confidence intervals and *p*-values with minimal mathematical distractions such as the "machinery of numerical approximations based on the normal distribution and its many subsidiary cogs" (para. 1). By not waiting to teach statistic inference after "the culmination of a long development of prerequisite material on sampling distributions, formulas for standard errors, standard reference distributions, central limit theorems, and formulas for standardizing values" (Lock et al., 2014, para. 1), simulation-based methods allow more emphasis on the interpretation and underlying logic of statistic inference, including the connection between randomness in study design and inferential reasoning. In particular, students can learn about the overall statistical process and have the tools to answer a genuine research question much earlier in the course (Roy et al., 2014). Throughout the course, instruction can focus repeatedly on the overall statistical process, rather than positioning data collection, data exploration, probability, and statistical inference as unrelated topics. Simulation-based inference can also make abstract concepts more concrete (Chance & Rossman, 2006), with the potential to make student thinking visible (Case & Jacobbe, 2018), enhancing student understanding as well as the instructor's ability to diagnose misunderstandings. Such changes to the content and sequencing of the material, now often dubbed "simulation-based inference" (SBI), keeps students "closer to the data," and naturally allows for more active learning and construction of knowledge in the classroom.

In the years since Cobb's (2007) recommendation, several new textbooks (see Section 2) have taken different paths for implementing his suggestions and "simulation-based inference methods are increasingly common in introductory statistics courses as a complement or substitute for theory-based inference" (Case & Jacobbe, 2018). Many of these ideas overlap with an overhaul of New Zealand high school curriculum and a "staged development path" to help students foster a more intuitive understanding of statistical inference (Wild et al., 2011, p. 247). Zieffler et al. (2008) contributed by describing the state of the field regarding informal statistical inference, using simulation as a key tool to illustrate informal statistical inference in the classroom. Lee et al. (2015) argued there are many positive aspects of the SBI approach, but also opportunities to improve and enhance that approach.

Formal classroom research evaluating these proposed curricular changes is still emerging. Preliminary assessment results, primarily at single institutions, have shown promising benefits to a simulation-based approach. For example, Maurer and Lock (2015) found advantages to using bootstrapping when introducing confidence intervals. Hildreth et al. (2018) saw improvements from using simulation-based inference curriculum on six key statistical concepts. Beckman et al. (2017) saw improved cognitive transfer outcomes when comparing a simulation-based curriculum to a traditional

approach. Pfannkuch and Budgett (2014) found that visual inference tools using bootstrapping and randomization tests facilitated the development of statistical inferential concepts, while Lane-Getaz (2017) and Reaburn (2014) found students better able to define and use *p*-values in a course using simulations. Three papers examining outcomes from the *Introduction to Statistical Investigations* (ISI) curriculum (Chance & McGaughey, 2014; Tintle et al., 2011, 2012) have found similar or improved pre- to post-course outcomes when using SBI curricula compared to prior consensus curricula (e.g., Advanced Placement Statistics; Roberts et al., 1999).

The literature to date has focused primarily on smaller samples from limited numbers of institutions, often comparing only two curricula. This has led some to call for additional studies to enhance generalizability, including measures of student demographics, measures of student ability, and measures of student attitudes (e.g., Hildreth et al., 2018). Furthermore, the unique pedagogical aspects, ordering, and focus of the International Statistics Institute (ISI) curriculum on the overarching statistical process (see Tintle et al., 2011, 2014 for additional details), combined with promising preliminary data (e.g., Tintle et al., 2011, 2012) lead to questions about whether unique aspects of the ISI curriculum facilitate different learning outcomes compared to other SBI curricula. To investigate these questions and others, our team led a multi-institution, three-year assessment project (2014–2017) with funding from the National Science Foundation (DUE-1323210). We collected data on student, instructor, and institutional characteristics, and both pre- and post-course outcomes, with an emphasis on students' conceptual understanding. In this project, we utilized an adapted form of the Comprehensive Assessment of Outcomes in Statistics (CAOS) instrument, a widely used and validated instrument focused on conceptual understanding (Section 3.4).

In this paper we address the following research questions:

1. Are there differences in pre/post-changes in conceptual understanding depending on the type of curricula used? In particular, do students with simulation-based inference curricula show similar gains to non-simulation-based curricula across course content areas? If not, what are areas of weakness that can be addressed through curricular revision? Are these areas different for the ISI curriculum?
2. How does textbook choice compare to other student-level (e.g., ACT score), instructor-level (including years of experience), or classroom-level characteristics (including student background, instructor experience) in terms of impact on student conceptual learning gains pre- to post- course?

Although textbook choice is likely a proxy for several variables and this multi-institutional observational study will not reveal the components (e.g., use of active learning, focus on modeling, more student-driven technology, teacher experience, instructional time) that are most impactful on student learning, we hope to gain insight across a variety of courses as to where introductory tertiary students are still struggling, and whether the gains and struggles differ across textbook choices and types of students.

2. SIMULATION-BASED INFERENCE TEXTBOOKS

The last few years have seen the development of several full curricula/textbooks for introductory algebra-based statistics courses that focus on simulation-based inference (SBI). In each case, instruction on inferential methods is preceded by simulations illustrating the reasoning behind the methods. Some of the key distinctions between these approaches include:

- When simulations are first used for inference.
- Whether or not bootstrapping is highlighted as a simulation method in addition to randomization tests.
- Technology used.
- Target audience.

These distinctions are summarized in Table 1. In the teacher survey, instructors were asked to indicate the textbook they had adopted, but over the course of the three years of data collection described here, different editions could have been used, so in the table we focus on the main title of the text and not the edition.

Table 1. Textbooks using simulation-based inference (SBI) in the introductory, algebra-based statistics course

Textbook	First use of SBI	Bootstrapping	Technology	Target audience
Introduction to Statistical Investigations (*ISI*) (Tintle et al., 2015)	Ch. 1.	No	Rossman/Chance applets	Algebra pre-req
Statistics: Unlocking the Power of Data (*Lock5*) (Lock et al., 2013)	Ch. 3.	Yes	*StatKey*	Algebra pre-req
Statistical Reasoning in Sports (Tabor & Franklin, 2013)	Ch. 1.	No	Graphing calculator	High school
CATALST Project's *Statistical Thinking: A simulation approach to modeling uncertainty* (Zieffler & Catalysts for Change, 2015)	Ch. 1.	Yes	*TinkerPlots*™	Algebra pre-req
Introductory Statistics with Randomization and Simulation (Diez et al., 2014)	Ch. 2.	Yes	*R*	Algebra pre-req

Several instructors also indicated they did not use a textbook or solely used their own materials that were hybrids of these approaches (e.g., Hildreth et al., 2018; Malone & Hooks, 2012). We did not evaluate the extent of simulation-based inference in these materials. Textbooks also varied by the extent of coverage of parametric methods. For example, the original CATALST curriculum did very little with parametric tests (two-sample *t*-tests), compared to the ISI curriculum, which included SBI methods before ANOVA, regression, and Chi-square tests. We did not ask instructors to indicate the final chapter they covered in the course.

3. METHOD

3.1. RECRUTING FACULTY/STUDENT PARTICIPANTS

From 2013–2016, we sent an open call to introductory statistics faculty, primarily in the United States, using email listservs (e.g., ASA Section on Statistics Education, Isolated Statisticians, SIGMAA on Statistics Education). Faculty were asked to give their students an instrument assessing students' conceptual understanding of statistics and attitudes towards statistics (see Section 3.4) through *SurveyMonkey* during the first and last weeks of the term. Most faculty offered some incentive to their students (e.g., credit on a homework assignment) for participating. The instrument was generally given outside of regular class time, though a few instructors did include the concept inventory as part of the final exam. Section-level response rates were above 90% on the pre-test (mean = 0.995, *SD* = 0.196) and above 80% on the post-test (mean = 0.833, *SD* = 0.368), but about 25% of sections showed a response rate below 73% on the post-test.

3.2. DATA CLEANING AND INCLUSION CRITERIA

Examples of the lengthy data cleaning tasks included tracking students who changed instructors after the pre-test and reconciling discrepancies in demographic data between pre and post administrations (e.g., students or instructors providing inconsistent responses for sex, self-reported GPAs, age). Responses that were similar were averaged and discrepant responses (e.g., pre-post GPA differed by more than 0.25) were dropped. 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") and GPAs larger than 4.0 were truncated (in later versions of the

instrument, we ask students whether their institution's GPA is reported on a 4- or 5-point scale to allow rescaling). The cleaned student dataset was merged with a cleaned version of the teacher inventory data. We removed students who did not take at least 10 minutes to complete the instrument, who answered fewer than 80% of the questions on either test, who did not take both pre and post instruments (about 50% each year), or who opted out of permitting us to use their data for research purposes. Remaining students were matched pre to post and duplicate records removed (keeping the more complete record). The dataset is available from the authors upon request.

For the analyses in this paper, we removed students in statistics courses with a calculus prerequisite (typically aimed at more mathematically inclined students than algebra-based course), as well as 793 high school students (see Roy & Mcdonnel, 2018) for preliminary analysis of the high school students. Lastly, students with an achievable gain of -1.1 and lower (e.g., 78% correct on the pre-test, 33% correct on the post-test) were removed prior to analysis (less than 0.2% of students each year).

3.3. STUDY PARTICIPANTS

After applying the inclusion criteria described in Section 3.2, the final sample size was 10,514 students, across 503 sections, 194 instructors, and 86 institutions for academic years 2014/15, 2015/16 and 2016/17 (many instructors/institutions participated in multiple terms). Table 2 summarizes key student and instructor characteristics noted from the demographic survey and the instructor survey. Note that 31% of the instructors were tenured, 20% tenure track, 21% full-time lecturer, and 27% graduate teaching assistants.

Table 2. Student and instructor characteristics

Students	No previous stat course	First generation student (y2, 3)	Taking course for major	% Female
	66.6%	24.4%	78.5%	65.0%
Instructors	≤ "very little" data experience	No knowledge of GAISE	Mean/Median years teaching statistics	% Female
	35.2%	35.9%	9.7/6	55.2%

Appendix Table A.4 summarizes the course prerequisite, type of department teaching the course (statistics or statistical sciences, mathematics, and other), Carnegie Classification, and student type as reported by instructors at each institution. For student type, most students were lower classmen (freshmen or sophomores), with similar breakdown as to whether the course aimed to be a (lower division or upper division) general education course vs. a required course in their academic discipline.

3.4. INSTRUMENTS

We used a 32-question multiple-choice concept inventory adapted from the Comprehensive Assessment of Outcomes in Statistics (CAOS) instrument from the University of Minnesota (delMas et al., 2007), which is a well-established assessment instrument for assessing outcomes of an introductory college statistics course. Rather than use an instrument focused only on inferential reasoning (e.g., Lane-Getaz, 2007; Tobias-Lara & Gomez-Blancarte, 2019) or one focused on more modern topics (e.g., Ziegler & Garfield, 2018), we wanted to first compare student performance on items designed for non-simulation-based courses. In validating and revising the CAOS instrument, delMas et al. (2007) asked statistics education experts and instructors to rate their agreement that the questions covered desired learning goals for any tertiary introductory statistics course. This validation process occurred before the emerging increase of textbooks centering around SBI. Topics on this inventory include data collection, simulation/probability (but not simulation-based inference), descriptive statistics, confidence intervals, significance tests, and scope of conclusions—content covered in both SBI and non-SBI courses. Our modifications to the CAOS instrument included revising some less frequently chosen distractors, reordering questions, and adding several new variations/items which covered questions related to the impact of sample size and common misinterpretations of the *p*-value (large *p*-value is evidence of the

null being true, confusing *p*-value with sample statistics). Appendix Table A.1 describes the questions on our instrument, including these eight new questions, with notes on the most similar CAOS questions. Appendix Table A.2 notes changes to the instrument made after Year 1. This concept inventory was combined with the SATS-36 instrument (Student Attitudes Toward Statistics; Schau, 2003) into one instrument.

As discussed in Tintle et al. (2018), the instrument showed good reliability (Cronbach's alpha > 0.65), construct validity (e.g., stronger item-total correlations post-course than pre-course), and good predictive validity (moderate correlations with external measures of quantitative understanding (e.g., ACT score, $r = 0.39$); associations with positive attitudes and post-test scores ($r > 0.25$ for 5 of 6 SATS subscales). Additional details on reliability and validity for a single year of this sample (2016-2017) are provided in Tintle et al (2018).

For analysis, the 36 questions on the concept inventory were regrouped into 24 question-sets, so sets of 2–3 questions with the same prompt and assessment goal (e.g., a series of valid/invalid statements of a confidence level interpretation) were scored together (see Appendix Table A.3). Each question and question set were coded as correct or incorrect (with partial credit on question sets), and the percentage correct was computed from the 24 possible points.

At the beginning of the course, instructors indicated the number of sections they taught, class sizes, and textbook choice. At the conclusion of the course, instructors were also asked to complete a survey ("teaching inventory") 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). The instructor questions were loosely based on Zieffler et al., 2012 (see also Fry, 2014; Parker et al., 2014).

4. RESULTS

4.1. GAINS IN CONCEPTUAL UNDERSTANDING

On the pre-test, students performed similarly across all three textbook classifications (Table 3), with most students answering between 35% and 55% of the 24 question sets correctly. Patterns were similar in the three years of data (details not shown).

Table 3. Student pre-test scores by textbook classification (Years 1-3)

	ISI	Other SBI	NonSBI
n	2872	3251	4139
Mean	0.477	0.482	0.471
SD	0.113	0.114	0.112

The primary response variable of interest was students' post-course performance on the concept inventory. To adjust for pre-test scores and possible ceiling effects, we also considered *achievable gain* = *(post – pre)/*(1 – *pre*) as a measure of student improvement (aka "single-student normalized gain", e.g., Colt et al., 2011; Hake, 1998). Table 4 shows the mean achievable gain scores across the textbooks for the three years. Although the achievable gains were modest, we saw a consistent pattern across the years. Notably, non-SBI curricula have the lowest achievable gain, followed by Other SBI curricula and the ISI curriculum consistently showing the largest achievable gain on average. The patterns held true after adjusting for student background data as discussed in Section 4.2. Also notable was the considerable within-section variability (Figure 1 for Year 3).

Table 4. Mean achievable gain scores for 2014/2015–2016/2017 school years, across textbook classifications

	Ach gain	Overall	ISI	Other SBI	NonSBI
Year 1	Mean	0.148	0.179	0.168	0.109
	SD	0.261	0.262	0.264	0.254
Year 2	Mean	0.117	0.178	0.140	0.077
	SD	0.250	0.248	0.259	0.239
Year 3	Mean	0.137	0.173	0.158	0.076
	SD	0.257	0.245	0.269	0.245

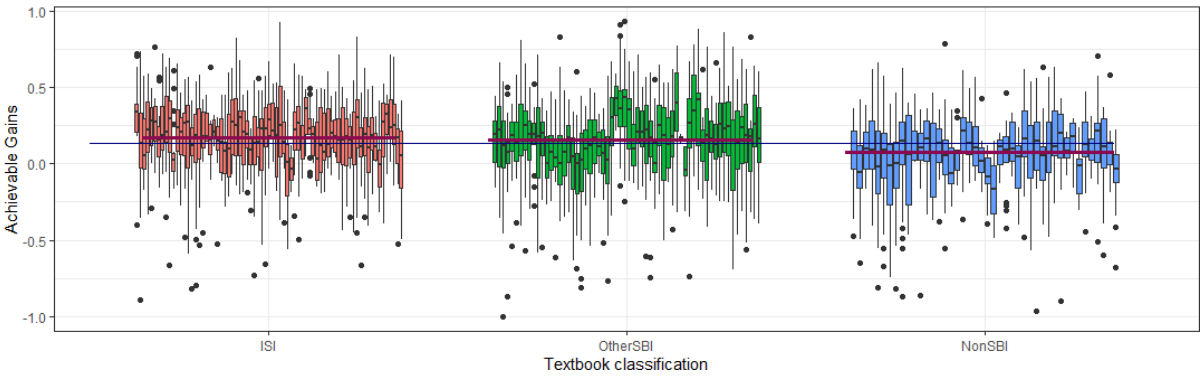

Figure 1. Boxplots of achievable gain by textbook for 179 sections (Year 3). Red lines indicate averages by textbook classification.

Table 5 compares the gains (*post – pre*) in the proportion correct for each subcategory of the concept scale by textbook classification. Overall, student gains were highest for the confidence intervals and tests of significance subscales. There was evidence of higher gains with the simulation-based approaches (ISI and Other SBI), particularly in questions about data collection, significance testing, and simulation, with the ISI curriculum outpacing Other SBI curricula on significance, and Other SBI outpacing ISI on confidence intervals. The three groups were more similar for questions about descriptive statistics.

Table 5. Comparison of pre-concept scores and gains in proportion correct overall by textbook across concept scale subcategories

Subcategory of conceptual inventory (# items)	Pre-test scores Mean (*SD*)	Mean Gain (*SD*)			
		Overall	ISI	Other SBI	NonSBI
Data collection (4)	0.538 (0.211)	0.057 (0.275)	0.090 (0.281)	0.090 (0.273)	0.012 (0.266)
Descriptive (7)	0.484 (0.205)	0.054 (0.222)	0.060 (0.221)	0.056 (0.221)	0.050 (0.222)
Confidence Int (5)	0.347 (0.174)	0.113 (0.258)	0.105 (0.253)	0.134 (0.262)	0.104 (0.257)
Significance (10)	0.559 (0.182)	0.101 (0.222)	0.143 (0.218)	0.102 (0.220)	0.070 (0.220)
Simulation (7)	0.405 (0.229)	0.070 (0.271)	0.109 (0.268)	0.090 (0.272)	0.028 (0.267)

Figure 2 compares the textbook categories on the individual questions, illustrating the variability in gains on individual questions within the same concept subcategory. Green lines indicate questions with higher average scores on the post-test compared to the pre-test. Table 6 highlights several interesting comparisons with a complete summary of results for all questions in the Appendix (Table A.3).

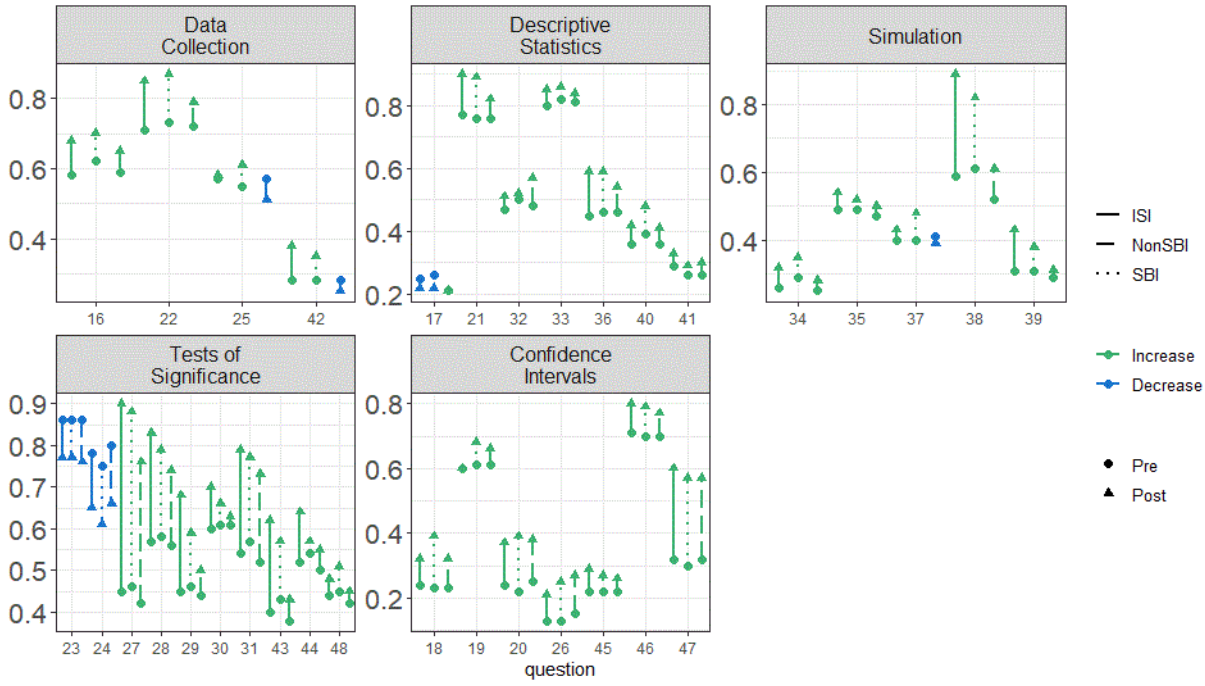

Figure 2. Question by question pre/post-test performance by textbook category (Triads correspond to ISI, SBI, NonSBI)

Table 6. Selected question by question comparisons on mean pre-test, post-test, and achievable gain scores

	ISI			Other SBI			NonSBI		
	Pre	Post	Ach Gain	Pre	Post	Ach Gain	Pre	Post	Ach Gain
Highest pre-test scores									
Q21: Comparing two distributions	0.77	0.90	0.56	0.76	0.89	0.53	0.76	0.82	0.26
Q33: Matching graph to variable description	0.80	0.85	0.25	0.82	0.86	0.23	0.81	0.84	0.14
Largest SBI improvement									
Q27: Recognizing the goal of a small *p*-value	0.45	0.90	0.84	0.46	0.88	0.78	0.42	0.76	0.60
Q28: Recognize *p*-value is not probability of null	0.57	0.83	0.59	0.58	0.79	0.51	0.56	0.74	0.42
Q31: Recognize *p*-value is not the statistic	0.54	0.79	0.54	0.57	0.77	0.46	0.52	0.73	0.43
Q29: Recognize valid interpretation of *p*-value	0.45	0.68	0.42	0.46	0.59	0.24	0.44	0.50	0.12
Q43: Inferential reasoning	0.40	0.62	0.36	0.43	0.58	0.25	0.38	0.42	0.08
Q46b: Recognize impact of confidence level on width	0.32	0.60	0.41	0.30	0.57	0.38	0.32	0.57	0.36
Q38: Recognizing correct simulation model	0.59	0.89	0.72	0.62	0.82	0.53	0.52	0.61	0.18
Largest NonSBI improvement									
Q32: Matching graph to description	0.47	0.51	0.08	0.50	0.52	0.02	0.48	0.57	0.17
Lowest improvement (all)									
Q23: Small sample size	0.88	0.77	-0.77	0.86	0.77	-0.64	0.86	0.76	-0.69
Q24: Large *p*-value is evidence for the null	0.78	0.65	-0.58	0.75	0.61	-0.60	0.80	0.66	-0.65
Q25: Association vs. causation	0.57	0.58	0.01	0.55	0.61	0.13	0.57	0.51	-0.13
Q17: Most appropriate graph	0.25	0.22	-0.04	0.26	0.22	-0.04	0.21	0.21	0.00
Q42: Purpose of random assignment in study design	0.28	0.38	0.14	0.28	0.36	0.11	0.28	0.25	-0.04
Poorest post-test scores									
Q26: Necessary sample size for US population	0.13	0.21	0.10	0.13	0.25	0.14	0.15	0.27	0.14

High pre-test scores (above 0.75)

- On the pre-test, students performed most strongly on Q21 and Q33, which dealt with comparing two dotplots of unequal sample size (one question with three options rather than CAOS' three valid/invalid statements) and selecting the histogram best matched to a variable description. (Our instrument did not include a CAOS question regarding linear associations as we have found students typically score highly on this question at the beginning of the course as well.)
- We also note from Figure 2 that students tended to score higher on pre-test Questions 23 and 24, but then had lower performance on the post-test (see below).

Largest improvements with SBI curricula (gain > 0.20)

- All curricula, but especially the SBI-based students, showed substantial improvement recognizing the purpose of a *p*-value (Q27). Similarly, all curricula showed substantial improvement in the ability to recognize *p*-value interpretations as the probability of the alternative hypothesis (Q28) and as the difference in conditional proportions (Q31) as incorrect.

- SBI curricula showed more improvement recognizing a correct interpretation of a *p*-value (Q29) and in making a conclusion from a *p*-value (Q43), with less growth by non-SBI students, on average.
- SBI curricula showed strong improvement in recognizing a correct simulation model (Q38), with little growth by non-SBI students on average, though many students (across textbook classifications) continued to consider other invalid "simulation models" (e.g., repeating the study) to be valid as well (Q37, Q39).
- Lastly, all curricula showed substaintial gains in student ability to recognize the impact of increasing the confidence level on the confidence interval width (Q46b).

More improvement with Non-SBI curricula than SBI

- The non-SBI students gained more on a question asking them to match a histogram to a variable description (Q32), but all three categories were around 50% pre and post. We also note the SBI curricula showed similar performance to non-SBI students on other descriptive statistics questions, a topic often assumed to be not emphaized in the new curricula.

Low improvement across all curricula (gains < 0.01)

- Across textbook categories, students tended to decrease in performance on questions asking about the role of sample size (Q23) and whether large *p*-values provide evidence *in favor of* the null hypothesis (Q24). This latter tendency is also shown in Q45, on which students indicate a value inside the confidence interval provides evidence *for* the null hypothesis at the beginning and the end of the course.
- Q17: This question asked students to select the most appropriate graph of a quantitative variable. At the beginning and end of the course, on average, students selected a case-value graph that looked bell-shaped, rather than one that best demonstrated the distribution of the variable.
- Students did not show much improvement on a question asking them to recognize that a causal conclusion could not be drawn from an observational study (Q25), though improvement was better with Other SBI students.
- SBI students gained more on average, though still not a lot, on a question asking about the purpose of random assignment (Q42), compared to a decrease for non-SBI students. However, we would have expected more improvement given the strong focus and reinforcement by those texts on the topic.

Poor post-course scores (post < 0.25)

- For Q26, students demonstrated poor performance on a question asking about the necessary sample size to adequately represent all 310 million U.S. residents, at the beginning and end of the course.

4.2. MULTILEVEL MODELS

Given the nested structure of our data, we used multi-level models to estimate the relative impact of textbook choice on student gains after accounting for a variety of covariates. There were some modest differences in the instrument in Year 1 (see Appendix A.2), so the primary modelling results were obtained by fitting models on data that pooled Years 2 and 3 (n = 7,536). Using a four-level model (institution, instructor, section, student), the intraclass correlation coefficient for the null model using achievable gain as the response is around 11% (similar for each year). By far, most of the variation is at the student level. Most of the remaining variation appeared to be at the institution level (6.6%) compared to the instructor level (3.5%) and section level (1%). Table 7 lists the variables, by level, used to build the multi-level models. Rather than using *achievable gain* in these models, we used *gain* including *pre-test score* and *pre-test score*^{2} as predictors.

Table 7. Variables using in model building

Student variables	Section variables	Instructor variables	Institution variables
• Pre-concept (quadratic) • Pre-attitudes (6) • Other SATS-36 questions (math competence, math performance, grade expectation, mastering confidence, stats career usage, major course, likelihood would choose, why took course) • GPA (quadratic) • Math SAT/ACT *z*-score (quadratic) • Age • Sex (Male or Female) • Previous stat course (2) • Area of study • First generation • Race (white, binary) • Status (year in school)	• Average pre-concept • Average pre-attitude (6) • Average overall attitude[d] • Section level GPA • Avg SAT/ACT *z*-score • Fall or Winter/Spring • Class size (start) • Response rate *Instructor reported* • Time of day (morning, midday, or after 2pm) • Length of class session • Math prereq • Incentive[a] for taking instrument, location, timing (codes) • Type of student (upper/lower/GE/reg) • Lecture type (primary delivery of new content) (2) • Percentage of class time in lecture	• Years of experience (overall intro stat) • Experience analyzing data • Position type • Type of advanced degree • Use of TA for leading section • GAISE familiarity • Attended ISI workshop • Sex (binary) • Textbook (ISI, OtherSBI, NonSBI)	• Dept type[b] • Schedule (# weeks)[c] • Carnegie classification • Average attitudes and overall GPA, SAT/ACT *z*-scores, pre-test, and response rate (11)

[a]Incentive offered for the post test was categorized as none, low stakes, high stakes (e.g, part of final exam). We also considered whether the post-test was given in or outside of class. Pre-test classifications were similar.
[b]One institution had multiple instructors from different departments.
[c]One institution had a year-long "dual-enrollment" course for college credit.
[d]Overall attitude was computed by averaging the SAT scales of affect, cognitive competence, interest, and value.

From the large list of variables in Table 7, we explored several model building strategies. Initially, we identified significant variables after using imputation (with the 'amelia' package v. 1.7.6 in *R*) on the variables with high rates of missingness (e.g., GPA, SAT/ACT *z*-score), without including the textbook variable. First, we systematically added in variables from different categories (e.g., attitude variables, other student-level variables, section-level variables, etc.), and then used backwards elimination to reduce the list of explanatory variables at each step. With this approach we identified 20 variables with *p*-values below 0.05 (using lmer, and Anova from the 'car' package v. 3.0-12 in *R*) and a model that explained 38% of the variation in gains (conditional R^2 from 'performance' package v. 0.8.0 in *R*). Second, we began with a model with all the variables and manually used backwards elimination down to 33 variables. These variables were combined with any additional variables from the initial set of 20, textbook, and some interactions of interest were added. These variables were then applied to the non-imputed data (n = 3,036). This model was then trimmed to 24 variables. Table 8 shows the first 15 terms, sorted by significance (using t and X^2 statistics). Apart from the interaction between student gender and section value pre, the largest VIF values (from performance package) fall below 7.

Table 8. Most significant variables from final multi-level model

Variable	Coefficient	*t*-statistic	X^2 statistic
Pre-test *z*-score	-0.073	-41.70	1739
GPA *z*-score	0.037	11.40	254
SAT/ACT *z*-score	0.018	7.71	112
Pre-test *z*-score^2	0.008	7.37	54.3
GPA *z*-score^2	0.010	6.88	50.3
Textbook classification			55.5
NonSBI	-0.047	-7.20	
Other SBI	-0.013	-2.126	
ISI (baseline)			
Value	0.015	5.74	32.9
Sex (Male)	0.179	2.45	26.5
Institution SAT/ACT	0.054	4.93	24.3
Mastering confidence	0.007	4.69	22.0
Institution overall attitude	-0.135	-4.05	16.4
Institution interest	0.092	4.03	16.3
First gen (yes)	-0.015	-3.88	15.0
Effort	-0.008	-3.56	12.7
Female × Section value	0.037	3.35	11.5

By far, the strongest predictors of *gain* on the 24-point concept inventory were the students' prior abilities and performance as measured by pre-test score, GPA, and SAT/ACT (accounting for more than 20% of the variation in student gains). The relationship with pre-test score was concave up: students with below average pre-test scores showed the strongest gains. The relationships with GPA and SAT/ACT were also concave up, showing increasingly higher gains for above average students. Overall, males and those who had confidence in their ability to master the material, showed higher gains.

After adjusting for these variables, textbook classification was the next strongest predictor, with non-SBI students showing lower gains on average compared to ISI students, and not a large difference between ISI and Other SBI students. The coefficients of the textbook classifications aligned with the earlier results in unadjusted analyses (e.g., Table 4). Similarly, using forward selection in a single-level model with the same predictors, textbook was the fourth variable to enter the model after pre-concept score, SAT/ACT, and GPA.

First generation students, and those who believed they would put a lot of effort into the course, on average exhibited lower gains. Surprisingly, several institution-level variables remained in this final model. Students at institutions with higher average SAT/ACT scores and higher aggregated interest in statistics, showed higher gains on average, but students at institutions with a higher overall attitude score, showed lower average gains. These characteristics appear more important to the model than Carnegie Classification.

Of the interactions with textbook classification, student sex, and instructor sex, the strongest appeared to be an interaction between section-level value of statistics and student sex. Figure 3 reflects this interaction, indicating that sections with higher average value of statistics more positively influenced students who identified as female rather than male. We can cautiously interpret that males are less impacted by this element of classroom culture, whereas females gain more when surrounded by peers that value the study of statistics. Another interesting but less significant interaction was GPA × textbook (not shown), which indicated that the slope of GPA was largest for the ISI instructors and smallest for the non-SBI instructors, resulting in a larger difference between textbooks for those students entering the course with above average GPAs.

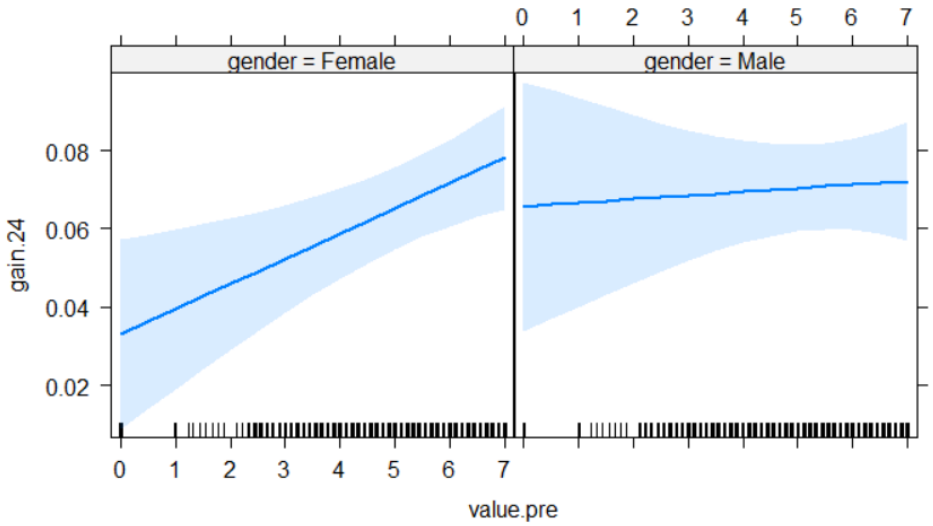

Figure 3. Interaction between student-provided sex and value of statistics

4.3. SENSITIVITY ANALYSES

We performed three sensitivity analyses (details not shown). First, to confirm that model results were robust by year, we fit the final model on Years 2 and 3 separately and estimated similar effects for all variables in the final model. We also noted the effects were similar when run with the imputed data and the non-imputed data. Finally, we used propensity score weighting to create more equivalent groups between instructors choosing SBI and non-SBI curricula. After creating these more equivalent groups, models comparing groups estimated similar sized differences in gain between curricula (Appendix Figure A.1).

We also refit the final model for only the students who scored below average on the pre-test. The textbook classification variable was still highly significant, with coefficients of -0.043 for NonSBI and -0.0065 for Other SBI compared to ISI. We also focused on the relationship between *gain* and the pre-test score for SBI students and non-SBI students. Figure 4 suggests that students with lower pre-test scores tend to see slightly higher gains with the SBI curricula, and similar average gains with students who scored above average on the pre-test.

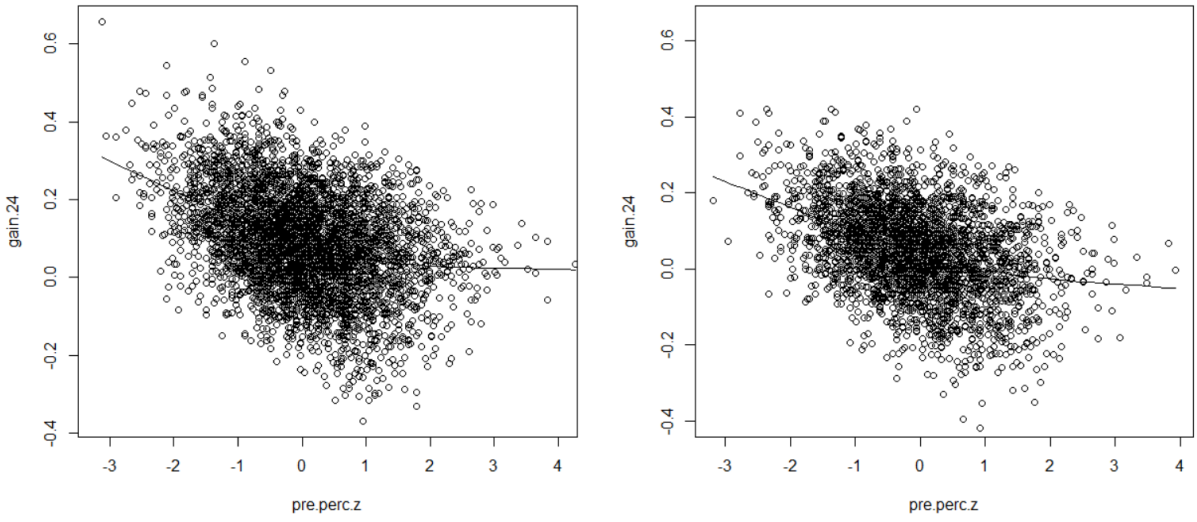

Figure 4. Smoother showing quadratic relationship with pre-concept scores for SBI (ISI and non-ISI) curricula and non-SBI curricula

5. DISCUSSION

After collecting data across multiple instructors and institutions for three years, we were able to take a deeper look at numerous variables potentially related to student learning in introductory statistics, as well as explore differences in student understanding arising from simulation-based and non-simulation-based textbooks. This dataset provided us the opportunity to update earlier reports (e.g., delMas et al., 2007), to expand upon more recent studies (e.g., Hildredth et al., 2018), and to include variables at several levels in the educational process, the latter allowing us to further explore what components of the student experience (e.g., section-level and institution-level) are and are not associated with students' conceptual learning in introductory statistics courses.

We continue to see trends similar to those observed in these earlier papers. For example, students' attitudes coming into a course appear related to their learning, including their feelings of cognitive competence and prior performance in mathematics courses. Students' attitudes, however, are not as effective predictors as their previous performance (e.g., GPA, SAT/ACT). In fact, students' GPA, SAT/ACT score, and prior understanding (pre-test score) were the three variables most predictive of students' gain in conceptual understanding and accounted for approximately 23% of the variation in students' gain. Of interest is the quadratic nature of these associations, demonstrating the potential for larger gains in students who enter a course with less background knowledge, perhaps more so for SBI courses. Although the strength and form of these associations are in many ways expected, this leaves instructors with little consolation as none of these three factors are modifiable by the instructor of introductory statistics. Still, conversations about the value of statistics early in the course may help contribute to student gains. Institutional culture may also play a role, though perhaps more for students who identify as female rather than male.

Although we cannot explain much of the instructor-to-instructor variation with our instructor level variables, we found that after adjusting for such instructor and institutional level effects in our multilevel model, the textbook choice effects were significant. In particular, our analysis points to *textbook choice* being the most important modifiable factor impacting students learning, with curricula using simulation-based inference (ISI and Other SBI) showing achievable gains of 14–18% compared to 7–11% for non-SBI curricula (Table 4). This improvement was primarily concentrated in concepts relating to data collection, significance, and simulation—consistent with theoretical arguments about the potential benefits of SBI (Cobb, 2007). In particular, student performance on Question 27, recognizing that small rather than large *p*-values establish strong evidence in favor of a research conjecture, with many sections having 100% correct responses on the post-test, suggested that SBI students gained better understanding of the use of *p*-values in tests of significance. ISI tended to yield slightly higher achievable gain than Other SBI curricula, results that continued to hold in multivariable analyses, which accounted for a wide variety of other student, section, instructor, and institutional level variables. We also found a few differences across the simulation-based curricula. In particular, there were distinct subscales (e.g., confidence intervals) and questions (e.g., Q25 on the inventory) for which Other SBI curricula outperformed ISI and vice versa. We note, however, although these results are promising, the ultimate difference, on average, from textbook choice, is equivalent to an additional 1-2 points on the 24 question-set inventory.

5.1. IMPLICATIONS FOR SBI TEACHING

Although SBI methods show similar student impact among first and long-time instructors (Chance et al., 2017), these assessment results indicate there are still numerous areas that need particular attention when implementing these methods. When asked to select the graph representing a histogram for a table of data (Q17), both SBI and non-SBI students tended to favor graphs that represented bell-shaped distributions rather than histograms. Introductory courses should focus more attention on identifying variables likely to follow a normal distribution and why. Furthermore, especially with the SBI approach, we encourage instructors to include graphs of other statistics (e.g., *F*-statistic, Chi-square statistic, Mean Absolute Difference) that do not follow a normal distribution and to emphasize to students that the variability in the null distribution is the key feature in inferential statistics. Although not assessed on our instrument, we also conjecture that use of SBI methods to supplement traditional instruction for "theory-based" approaches such as *t*-tests and Chi-square tests can improve student

understanding of the appropriate role of *p*-values and confidence intervals in statistical analyses, which may help students understand the limitations of statistical inference rather than rejecting its use.

The results for Questions 37–39 were a bit discouraging because even students in a simulation-based course could not correctly identify appropriate simulation strategies from situations that were not "simulations" (e.g., repeating the actual study). We recommend careful attention to the role of modelling (e.g., Garfield et al., 2012) and how the simulation model differs from the data production process. One suggestion is using more examples in which students are directly involved in the data collection process to help students differentiate observed data and simulated data. We have found that students also easily confuse "number of samples" with "sample size," and the results for Q26 indicated that students appeared to need additional reinforcement on the role of sample size (and population size) in statistical analyses (though this was found to be an area of weakness across all textbook categories).

We also note that all textbook groups had decreased performance from pre- to post-test on Q24, indicating that a large *p*-value can provide evidence *for* the null hypothesis. Instructors need to repeatedly remind students of the distinctions in these conclusions. Similarly, Q45 revealed a need for more discussion on the duality of confidence intervals and tests of significance, so students don't see them as separate procedures/providing additional evidence, but rather two different ways of summarizing a study each with its pros/cons.

5.2. FUTURE DIRECTIONS

More research is needed to identify the components of these novel curricula (e.g., focus on active learning, pedagogical knowledge of instructors, connection to the statistical investigation process) as they are most helpful in achieving these modest gains. Smaller, more focused studies could use randomization and/or cross-over designs to better illustrate the impact of lesson content and learning and teaching styles (e.g., Maurer & Lock, 2015). Focus-groups and think-aloud-protocols could be used to better understand how and when students make connections. Furthermore, we note the actual impact of SBI curricula on overall student understanding is relatively small. Better understanding of the reasons why SBI curricula are making some difference may also help point to areas of further innovation in SBI to increase the impact and/or suggest ways to translate the impact to other domains that showed less evidence of a difference (e.g., confidence intervals, descriptive statistics). Importantly, such experiments should include students from diverse backgrounds and institutions to ensure widespread generalizability, and to evaluate whether pedagogical best practices are similar across students and institutions. Further research is also needed to unpack how much simulation-based inference is necessary to reinforce student reasoning and whether the timing and sequencing of such discussions impact student performance.

Additionally, recent evidence at single institutions (Tintle et al., 2018, see also Figure 4) has suggested that students entering the course with weaker quantitative backgrounds may be among those that benefit the most from SBI curricula. Further research is needed to explore this trend across a wide range of institutions to confirm the generalizability of these findings and potentially differentiate best practices curricularly or pedagogically when working with these groups. Further research can also unpack differential relationships with attitudes vs. achievement (e.g., van Es & Weaver, 2018).

We also need to acknowledge that, even with the numerous variables considered in our anlayses, our models only explained approximately 1/3 of the total variation in students' gains. Although observational studies cannot be expected to generate the same level of control over variation as controlled experiments, it appears that there are still numerous, likely measurable factors that should be included in future studies. Future efforts could include utilizing richer instructor surveys, gathering information on student conceptual understanding from throughout the course of the semester, moving from self-reported to institutionally-reported student characteristics, and assessing students in a more controlled (e.g., monitored computer lab) environment. We do note in our sample a small amount of student data was obtained under proctored settings, which did show slightly higher gain scores, suggesting that our estimates of achievable gain may be under-estimated overall.

This does not mean that other changes to pedagogy or the student experience (e.g., class size), although not ranking in the top of our list of most important variables, are not worthwhile. For example, Posner (2014) suggested thinking carefully about the first day of class and the impact it can have on

student attitudes may still be "worth it," even though the ulimate benefit on students' understanding may be fairly small.

Another key area of future research is exploring retention after the end of the semester. Notably, our work here only looks at student learning gains at the end of the semester. To date, few studies have looked beyond that point, though preliminary evidence again suggests SBI may have benefits (Tintle et al., 2012). Further work is needed to explore whether and how learning gains from SBI curricula may be differentiated in the months and years after the course, and whether the learning gains may be related to other modifiable or non-modifiable student, instructor, or institutional characteristics. Little is currently known about retention more than four months after the course ends, and how this may be impacted by choices in the first statistics course, or student choices about subsequent courses and their timing (A second course in statistics? A course in a student's major that uses statistics?).

Finally, we acknowledge two limitations of this analysis. First, the sample here were all volunteers: instructors opted to participate. Although we have demonstrated the large variability and diversity in the sample (institutional, first generation, major, sex, previous experience with statistics, pre-test scores, instructor pedagogy, textbook choice, etc.), it is not a true random sample. Second, the instruments used here were designed for "traditional" introductory statistics courses and, although still focused on conceptual understanding, often fail to capture some of the big, important, and novel ideas that students leave SBI courses understanding (e.g., impact of test statistic choice; the overarching process of drawing conclusions from data). New instruments (e.g., Ziegler & Garfield, 2018) are needed that fully capture all areas of important student learning in modern introductory statistics courses.

6. CONCLUSIONS

Despite continued rapid gains in student enrollments and simultaneous advances in our understanding of best practices in statistics education, the vast majority of students do not show very large overall gains in conceptual understanding in introductory statistics courses, though they do on some individual questions more than others. A bleaker picture is painted when we realize that factors like student background and prior mathematical and statistical understanding continue to play a large part in what our students take from our introductory statistics courses. Although there are no "magic wands," our analysis provides the strongest evidence yet that, across widely varying student groups, instructional experience/pedagogy and institutions, incorporating a simulation-based inference curriculum may be an important, impactful, and "easy" aspect of learning and teaching that can be addressed for students. Despite this evidence, much additional research is needed, including improved assessment instruments that are adapted to emerging conceptual goals related to simulation-based inference ideas, better documentation of teaching practices, and controlled experiments to elucidate causality and determine best practices more directly. Such research should provide better understanding of the potential long-term benefits of SBI and other instructional choices. Furthermore, and very importantly, is the need to understand fully whether preliminary evidence of similar or improved learning gains from SBI courses in disadvantaged student groups and implied best practices are generalizable across institutions.

6.1. ACKNOWLEDGEMENTS

We thank the National Science Foundation for two grants (DUE-1140629 and DUE-1323210) that supported this research and the student researchers who contributed to these analyses. Thank you also to the editor, Jennifer Kaplan, and associate editors and reviewers for their feedback and patience.

6.2. CONFLICT OF INTEREST

Two of the authors of this manuscript (Nathan Tintle and Beth Chance) are also co-authors of the ISI curriculum development team. Co-authors were undergraduate Frost Research scholars at California Polytechnic State University.

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., & Witmer, J. (2005). *Guidelines for Assessment and Instruction in Statistics Education College Report*. American Statistical Association. https://www.amstat.org/docs/default-source/amstat-documents/2005gaisecollege_full.pdf

Beckman, M., delMas, R., & Garfield, J. (2017). Cognitive transfer outcomes for a simulation-based introductory statistics curriculum. *Statistics Education Research Journal*, *16*(2), 419–440. https://doi.org/10.52041/serj.v16i2

Case, C., & Jacobbe, T. (2018). A framework to characterize student difficulties in learning information from a simulation-based approach. *Statistics Education Research Journal*, *17*(2), 9–29. https://doi.org/10.52041/serj.v17i2

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 (ICOTS-9), Flagstaff, Arizona. International Statistical Institute. http://icots.info/icots/9/proceedings/pdfs/ICOTS9_6B1_CHANCE.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), Salvador, Brazil. International Statistical Institute. www.ime.usp.br/~abe/ICOTS7/Proceedings/PDFs/InvitedPapers/7E1_CHAN.pdf

Chance, B., Wong, J., & Tintle, N. (2017). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education*, *24*(3), 114–126. https://doi.org/10.1080/10691898.2016.1223529

Cobb, G. W. (1992). Teaching statistics. In L. A. Steen (Ed.), *Heeding the call for change: Suggestions for curriculum action* (pp. 3–43). Mathematical Association of America.

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

Colt, G. C., Davoudi, M., Murgu, S., & Zamanian Rohani, N. (2011). Measuring learning gain during a one-day introductory bronchoscopy course. *Surgical Endoscopy*, *25*, 207–216.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58. https://doi.org/10.52041/serj.v6i2.483

Diez, D., Barr, C., & Cetinkeya-Rundel, M. (2014). *Introductory statistics with randomization and simulation*. Openintro.org. https://openintro.org/book/isrs/

Doerr, H., & English, L. (2003). A modeling perspective on students' mathematical reasoning about data. *Journal for Research in Mathematics Education*, *34*(2), 110–136. https://doi.org/10.2307/30034902

Fry, E. B. (2014). Introductory statistics instructors' practices and beliefs regarding technology and pedagogy. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education.* Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9), Flagstaff, Arizona. International Statistical Institute. https://iase-icots/9/proceedings/pdfs/ICOTS9_C202_FRY.pdf?1405041868

Garfield, J. (1995). How students learn statistics. *International Statistical Review*, *63*(1), 25–34.

Garfield, J. & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review, 75*(3), 372–396.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM Mathematics Education*, *44*, 883–898. https://doi.org/10.1007/s11858-012-0447-5

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education*, *10*(2). http://dx.doi.org/10.1080/10691898.2002.11910665

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.

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), 103–120. https://doi.org/10.52041/serj.v17i1.178

Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modeling them. *International Journal of Computers for Mathematical Learning*, *12*, 217–230. https://link.springer.com/article/10.1007/s10758-007-9123-1

Lane-Getaz, S. J. (2007). Toward the development and validation of the reasoning about *p*-values and statistical significance scale. In B. Philips & L. Weldon (Eds.), *Proceedings of the ISI/IASE Satellite Conference on Assessing Student Learning in Statistics*. International Statistical Institute. http://www.stat.auckland.ac.nz/~iase/publications/sat07/Lane-Getaz.pdf

Lane-Getaz, S. J. (2017). Is the *p*-value really dead? Assessing inference learning outcomes for social science students in an introductory statistics course. *Statistics Education Research Journal*, *16*(1), 357–399. https://doi.org/10.52041/serj.v16i1.235

Lee, H. S., Doerr, H. M., Tran, D., & Lovett, J. N. (2015). The role of probability in developing learners' models of simulation approaches to inference. *Statistics Education Research Journal*, *15*(2), 216–238. https://doi.org/10.52041/serj.v15i2.249

Lock, R., Frazer Lock, P., Lock Morgan, K., Lock, E., & Lock, D. (2013). *Statistics: Unlocking the power of data* (First edition). Wiley.

Lock, R., Frazer Lock, P. F., Lock Morgan, K., Lock, E., & Lock, D. (2014). Intuitive introduction to the important ideas of inference. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education.* Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9), Flagstaff, Arizona. International Statistical Institute. http://icots.info/icots/9/proceedings/pdfs/ICOTS9_4A3_LOCK.pdf

Malone, C., & Hooks, T. (2012, July 31). *Finding an appropriate balance between simulation-based and traditional methods in the teaching of statistical inference* [Paper presentation]. Statistics: Growing to Serve a Data-dependent Society, Joint Statistical Meetings, San Diego.

Maurer, K., & Lock, E. (2015). Bootstrapping in the introductory statistics curriculum. *Technology Innovations in Statistics Education*, *9*(1). https://doi.org/10.5070/T591026161

Moore, D. M. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*(2), 123–165.

Parker, N., Fry, E., Garfield, J., & Zieffler, A. (2014). Graduate teaching assistants' beliefs, practices, and preparation for teaching introductory statistics. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education.* Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9), Flagstaff, Arizona. International Statistical Institute. https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_C200_PARKER.pdf?1405041867

Pfannkuch, M., & Budgett, S. (2014). Constructing inferential concepts through bootstrap and randomization-test simulation: A case study. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Education.* Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9), Flagstaff, Arizona. International Statistical Institute. http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8J1_PFANNKUCH.pdf

Posner, M. (2014). A fallacy in student attitudes research: The impact of the first class. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Education.* Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9), Flagstaff, Arizona. International Statistical Institute. http://icots.info/9/proceedings/pdfs/ICOTS9_1F3_POSNER.pdf

Reaburn, R. (2014). Introductory statistics course tertiary students' understanding of *p*-values. *Statistics Education Research Journal*, *13*(1), 53–65. https://doi.org/10.52041/serj.v13i1.298

Roberts, R., Scheaffer, R., & Watkins, A. (1999). Advanced Placement Statistics: Past, present, and future. *The American Statistician*, *53*(4), 307–320.

Roy, S., & Mcdonnel, T. (2018). Assessing simulation-based inference in secondary schools. Unpublished manuscript. http://www.isi-stats.com/isi/presentations/ICOTS2018-5.pdf

Roy, S., Rossman, A., Chance, B., Cobb, G., VanderStoep, J., Tintle, T., & Swanson, T. (2014). Using simulation/randomization to introduce *p*-value in Week 1. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Education.* Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9), Flagstaff, Arizona. International Statistical Institute. https://icots.info/9/proceedings/pdfs/ICOTS9_4A2_ROY.pdf

Saldanha, L. A., & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257–270.
Schau, C. (2003). Survey of Attitudes Toward Statistics (SATS-36). http://evaluationandstatistics.com/
Tabor, J., & Franklin, C. (2013). *Statistical reasoning in sports*. W.H. Freeman and Company.
Tintle, N. L., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2015). *Introduction to statistical investigations I* (First edition). 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://www.tandfonline.com/doi/full/10.1080/10691898.2018.1473061
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. https://doi.org/10.52041/serj.v11i1.340
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
Tobias-Lara, M. G., & Gomez-Blancarte, A. L. (2019). Assessment of informal and formal inferential reasoning: A critical research review. *Statistics Education Research Journal*, *18*(1), 8–25. https://doi.org/10.52041/serj.v18i1.147
van Es, C., & Weaver, M. (2018). Race, sex, and their influences on introductory statistics education. *Journal of Statistics Education*, *26*(1), 48–54. https://www.tandfonline.com/doi/full/10.1080/10691898.2018.1434426
Wild, C. J., Pfannkuch, M., Regan, M., & Horton, N. J. (2011). Towards more accessible conceptions of statistical inference. *Journal of the Royal Statistical Society: Series A* (Statistics in Society), *174*(2), 247–295. https://doi.org/10.1111/j.1467-985X.2010.00678.x
Zieffler, A., & Catalysts for Change. (2015). *Statistical thinking: A simulation approach to modeling uncertainty* (3rd ed.). Catalyst Press.
Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. *Journal of Statistics Education*, *16*(2). http://dx.doi.org/10.1080/10691898.2008.11889566
Zieffler, A., Park, J., Garfield, J. delMas, R., & Bjornsdottier, A. (2012). The Statistics Teaching Inventory: A survey on statistics teachers' classroom practices and beliefs. *Journal of Statistics Education*, *20*(1). https://doi.org/10.1080/10691898.2012.11889632
Ziegler, L., & Garfield, J. (2018). Developing a statistical literacy assessment for the modern introductory statistics course. *Statistics Education Research Journal, 17*(2), 161–178. https://doi.org/10.52041/serj.v17i2.164

BETH CHANCE
California Polytechnic State University
Department of Statistics
1 Grand Ave.
San Luis Obispo, CA 93407

APPENDIX

Table A.1. Mapping of concept inventory questions to most similar CAOS questions

#	Description	CAOS	Modifications
16	Identify relevant considerations in generalizing	38	Changed single forced choice check all that apply
22	Can we draw causal conclusion?	24	Changed context
25	Association vs. causation	22	Minor changes to answer choices
42	Primary purpose of random assignment in study design	7	Reduced number of answer choices
Descriptive statistics			
17	Identify which graph best represents distribution	6	Changed context
21	Comparing two distributions	11–13	Merged into one question
32	Matching graph to variable description	5	Minor changes to wording
33	Matching graph to variable description	3	Minor changes to wording
36	Comparison of conditional proportions	36	
40	Which histogram has the least variability/*SD*	14	Minor changes to wording
41	Which histogram has the greatest variability/*SD*	15	Minor changes to wording
Simulation/Sampling variability			
34	Larger sample sizes give less variable statistics	16	Changed context, wording
35	Which graph of statistics is most plausible	17	Gave dot plots rather than numbers
37	Valid/invalid design of simulation (repeat the study)	37	Single forced choice to 3 valid/invalid statements
38	Valid/invalid design of simulation (correct)	37	
39	Valid/invalid design of simulation (repeat the study)	37	
Tests of significance			
23	Could small sample size explain insignificance	23	Changed to valid/invalid statement
24	Is insignificant difference evidence in favor of null	New	Variation of CAOS 24
27	Is researcher hoping for small *p*-value or large *p*-value	19	
28	Valid/Invalid interpretation of *p*-value (probability null)	26	Lengthened statement in context
29	Valid/Invalid interpretation of *p*-value (valid)	25	
30	Valid/Invalid interpretation of *p*-value (probability alt)	27	Lengthened statement in context
31	Valid/Invalid interpretation of *p*-value (statistic)	New	
43	Inferential reasoning	New	
44	Impact of sample size on informal inference	New	
47	Which pair of dot plots have strongest evidence	New	
Confidence intervals			
18	Interpretation of confidence interval (individual)	28	Modified wording
19	Interpretation of confidence interval (valid)	31	Modified wording
20	Interpretation of confidence interval (statistic)	30	Modified wording
26	Sample size necessary for MOE for US population	New	
45	Duality between interval and conclusion	New	
46a	Impact of sample size on confidence interval width	New	
46b	Impact of confidence level on interval width	New	

CAOS questions not used

Q1: Pick best verbal narrative of descriptive output
Q2: Matching boxplot to histogram
Q4: One matching of histogram to variable description
Q8-10: Interpreting boxplots (used within curriculum)
Q18: Informal inference comparing two lists of outcomes
Q20: Interpretation of scatterplot (historically large pre-test percentage correct)
Q21: Impact of influential observation on judgement of linearity
Q29: Incorrect interpretation of confidence interval for 95% of observations
Q32: Recognize need for use of standard error vs. standard deviation
Q33: Match histogram to descriptive statistics
Q34-35: Match graph to sample/sampling distribution
Q39: Recognizing inappropriate extrapolation from regression line
Q40: Recognize possibility of Type I error

Table A.2. Modifications to questions after Year 1

Item	Modification/Impact
16	We formed one question from three valid/invalid questions to a single "check all that apply" question across four statements. Scores were scaled each year as the fraction of correct responses out of the 3 or 4 prompts. This change did impact the scores, lowering the percentage correct by 10-15 percentage points (pre and post).
42	For this question about the purpose of random assignment (generalization, causation, equal sample sizes), the last option was changed to "both a and b," making it much more attractive and lowering the overall percentage correct by roughly 20 percentage points (pre and post).
47	This question asks which pairs of dot plots showed stronger evidence of a treatment effect (#47). One graph option was changed so the correct answer required consideration of the within group variation in addition to the between group variation. This changed lowered scores by about 30 percentage points.
40, 41	For two questions judging least/greatest variation from histograms, the text "as measured by the standard deviation" was added, but this did not strongly impact scores.
24	The wording on this question, asking whether a large *p*-value provides evidence of X does not cause Y, was changed to whether a large *p*-value provides evidence that X does not impact Y, leading to a roughly 10 percentage point increase between the years (but not this is a question that saw a decrease in performance pre to post each year). Curiously, performance on Question 25 was noticeably lower in Year 1 (recognizing whether a causal conclusion could be drawn from an observational study), even though no change was made to the question.

Table A.3. Summary of concept inventory

		ISI				Other SBI				NonSBI			
#	Description	Pre	Post	Gain	A Gain	Pre	Post	Gain	A Gain	Pre	Post	Gain	A Gain
Data collection/Scope of conclusions													
16	Identify relevant considerations in generalizing from sample	0.58	0.68	0.10	0.25	0.62	0.70	0.08	0.22	0.59	0.65	0.06	0.15
22	Comparing two dot plots, can we draw causal conclusion?	0.71	0.85	0.14	0.49	0.73	0.87	0.14	0.52	0.72	0.79	0.07	0.25
25	Association vs. causation	0.57	0.58	0.01	0.02	0.55	0.61	0.06	0.13	0.57	0.51	-0.06	-0.13
42	Primary purpose of random assignment in study design	0.28	0.38	0.10	0.14	0.28	0.36	0.08	0.11	0.28	0.25	-0.03	-0.04
Descriptive statistics													
17	Identify which graph best represents distribution	0.25	0.22	-0.03	-0.04	0.26	0.22	-0.04	-0.04	0.21	0.21	0.00	0.00
21	Comparing two distributions	0.77	0.90	0.13	0.56	0.76	0.89	0.13	0.53	0.76	0.82	0.06	0.26
32	Set3: Matching graph to variable description	0.47	0.51	0.04	0.08	0.51	0.52	0.01	0.02	0.48	0.57	0.09	0.17
33	Set3: Matching graph to variable description	0.80	0.85	0.05	0.25	0.82	0.86	0.04	0.23	0.81	0.84	0.03	0.14
36	Comparison of conditional proportions	0.45	0.59	0.14	0.25	0.46	0.59	0.13	0.23	0.46	0.54	0.08	0.14
40	Set5: Which histogram has the least variability	0.36	0.42	0.06	0.10	0.39	0.48	0.09	0.15	0.36	0.41	0.05	0.09
41	Set5: Which histogram has the greatest variability	0.29	0.33	0.04	0.05	0.26	0.29	0.03	0.04	0.26	0.30	0.04	0.05
Simulation/Sampling variability													
34	Larger sample sizes give less variable statistics	0.26	0.32	0.06	0.08	0.29	0.36	0.07	0.10	0.25	0.28	0.03	0.03
35	Which graph of statistics is most plausible	0.49	0.54	0.05	0.09	0.49	0.51	0.02	0.04	0.47	0.50	0.03	0.05
37	Set4: Valid/invalid design of simulation (repeat the study)	0.40	0.43	0.03	0.05	0.4	0.48	0.08	0.14	0.41	0.39	-0.02	-0.03
38	Set4: Valid/invalid design of simulation (correct)	0.59	0.89	0.30	0.72	0.61	0.82	0.21	0.54	0.52	0.61	0.09	0.18
39	Set4: Valid/invalid design of simulation (repeat the study)	0.31	0.43	0.12	0.17	0.31	0.38	0.07	0.10	0.30	0.31	0.01	0.02
Tests of significance													
23	Could small sample size explain insignificant difference	0.86	0.77	-0.09	-0.77	0.86	0.77	-0.09	-0.64	0.86	0.76	-0.10	-0.69
24	Is insignificant difference evidence in favor of null	0.78	0.65	-0.13	-0.58	0.75	0.61	-0.14	-0.60	0.8	0.66	-0.14	-0.65
27	Is researcher hoping for small *p*-value or large *p*-value	0.45	0.90	0.45	0.82	0.46	0.88	0.42	0.78	0.42	0.76	0.34	0.60
28	Set2: Valid/Invalid interpretation of *p*-value (probability null)	0.57	0.83	0.26	0.59	0.58	0.79	0.21	0.51	0.56	0.74	0.18	0.42
29	Set2: Valid/Invalid interpretation of *p*-value (valid)	0.45	0.68	0.23	0.42	0.46	0.59	0.13	0.24	0.44	0.50	0.06	0.12
30	Set2: Valid/Invalid interpretation of *p*-value (probability alt)	0.60	0.70	0.10	0.25	0.61	0.66	0.05	0.14	0.61	0.63	0.02	0.03
31	Set2: Valid/Invalid interpretation of *p*-value (statistic)	0.54	0.79	0.25	0.54	0.57	0.77	0.2	0.46	0.52	0.73	0.21	0.43
43	Inferential reasoning	0.40	0.62	0.22	0.36	0.43	0.58	0.15	0.26	0.38	0.42	0.04	0.06

44	Impact of sample size on informal inference	0.52	0.64	0.12	0.24	0.54	0.57	0.03	0.05	0.50	0.55	0.05	0.10
47	Which pair of dot plots have strongest evidence	0.44	0.48	0.04	0.06	0.45	0.51	0.06	0.12	0.42	0.45	0.03	0.05
Confidence intervals													
18	Set1: Interpretation of confidence interval (prediction)	0.24	0.32	0.08	0.11	0.23	0.39	0.16	0.21	0.23	0.32	0.09	0.12
19	Set1: Interpretation of confidence interval (valid)	0.60	0.60	0.00	0.00	0.61	0.68	0.07	0.20	0.61	0.66	0.05	0.14
20	Set1: Interpretation of confidence interval (statistic)	0.24	0.37	0.13	0.18	0.23	0.39	0.16	0.21	0.25	0.38	0.13	0.17
26	Sample size necessary for MOE for US population	0.13	0.21	0.08	0.10	0.13	0.25	0.12	0.14	0.15	0.27	0.12	0.14
45	Duality between interval and conclusion	0.22	0.29	0.07	0.10	0.22	0.27	0.05	0.07	0.23	0.25	0.02	0.04
46a	Impact of sample size on confidence interval width	0.71	0.80	0.09	0.31	0.70	0.79	0.09	0.29	0.70	0.76	0.06	0.20
46b	Impact of confidence level on confidence interval width	0.32	0.60	0.28	0.41	0.30	0.57	0.27	0.38	0.32	0.57	0.25	0.36

Table A.4. Section/Institution characteristics (Percentages)

	Year 1	Year 2	Year 3
Prerequisite			
None	17	20	15
HS algebra	47	35	37
College algebra	29	34	37
Pre-calculus	0	2	1
Other (e.g., Math 101 or placement exam)	7	10	9
Type of department			
Statistics	29	32	30
Mathematics	63	61	60
Other	9	7	12
Carnegie classification			
Two-year college	14	16	14
Bachelor's	46	32	30
Master's	20	25	33
Doctoral university	20	27	23
Student type			
Lower division GE	---	26	53
Lower division required	---	53	27
Upper division GE/required	---	14/6	18/2

Note: The Student type question was not asked in Year 1.

Propensity Score Weighting

The 'cbps' package in *R* was used to create more equivalent groups between the instructors choosing SBI-focused and those not choosing SBI-focused textbooks. Instructors were matched based on *GAISE familiarity* (complete/mostly vs. some/no/unsure), *Carnegie classification*, *instructor sex*, *whether the instructor had taken additional statistics courses*, and the instructor-level averages for *pre-concept score* (quadratic) and *overall student attitudes score coming into the course*. After removing instructors with incomplete observations on these variables, we achieved covariate balance as illustrated in Figure A.1.

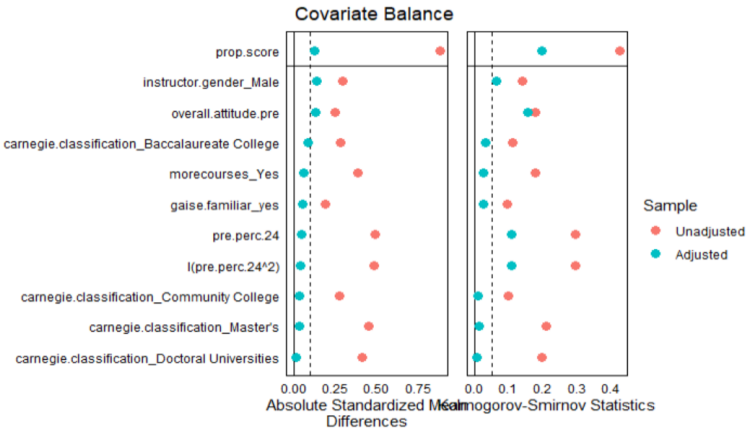

Figure A.1. Covariate balance from propensity score weighting on instructor-level variables

A model estimating the textbook effect on achievable gain, while still including adjustments for *pre-affect* and *Carnegie classification* illustrated an impact of 3.79 percentage points, in line with estimates in the primary multilevel analyses.

Term	Coefficient	*p*-value
Textbook (NonSBI)	-3.79	< .00001
Affect at baseline	1.96	.04
Carnegie classification – 2YC	-3.00	.008
Carnegie classification – Masters	-2.68	< .00001
Carnegie classification- PhD	-2.61	.003