ASSESSING THE ASSOCIATION BETWEEN QUANTITATIVE MATURITY AND STUDENT PERFORMANCE IN AN INTRODUCTORY STATISTICS CLASS: SIMULATION-BASED VS NON SIMULATION-BASED

Jill L. VanderStoep[1], Olivia Couch[2], Cassandra Lenderink[2]
[1]Department of Mathematics, Hope College, Holland, Michigan 49423
[2]Department of Mathematics and Statistics, Dordt College, Sioux Center, Iowa 51250
vanderstoepj@hope.edu

The use of simulation-based inference (SBI) methods when teaching introductory statistics continues to grow in popularity with evidence showing improvement in students' statistical thinking compared to theory-based (consensus) methods. Recent findings from two institutions comparing consensus curricula to SBI show that regardless of the measure used for a student's mathematical competency, SBI curriculum meaningfully impacted student learning. Here we revisit measures of mathematical competency with over 5,000 students at 81 institutions. In this larger sample, regardless of curriculum, weaker students tend to improve more by the end of the course than stronger students, however, SBI curricula outperformed consensus curricula for all levels of student mathematical competencies, with largest improvements seen on tests of significance and study design.

BACKGROUND

Since the late 1990s the United States, and countries around the world, have had a 'consensus curriculum' for teaching introductory statistics to students who have not necessarily had Calculus (Stat 101) (Schaeffer 1997). This consensus curriculum starts with descriptive statistics, moves into design, probability and sampling distributions and ends with statistical inference. Combined with the pedagogically focused GAISE guidelines (GAISE, Carver et.al., 2016), consensus instructors have worked on both the content and pedagogy of the Stat 101 course. However, for the past ten years there have been renewed calls to substantially reconsider the content in the first statistics course to focus on so-called 'simulation-based inference' methods including simulation, bootstrapping and permutation tests to help improve student learning and engagement (Cobb 2007).

In recent years, numerous curricula have now been developed which utilize simulation-based inference methods (e.g., Tintle et al. 2016; Lock et al. 2015). Preliminary evidence suggests small but statistically significant gains on normed tests of student performance (Tintle et al. 2011, Chance et al. 2017), but with medium sized effects in areas of particular focus of simulation-based inference curricula including tests of significance and study design. Recently, we demonstrated that, at two institutions, student improvement in statistically thinking as measured by the CAOS (delMas et al. 2007) test was associated with prior measures of student abilities, providing suggestive evidence that the impact of simulation-based inference was felt across various subgroups of students (Tintle et al. 2018). However, this analysis only focused on two institutions comparing a preliminary version of an SBI curriculum to the consensus curriculum. Here we expand this initial analysis to consider 85 separate institutions using a mix of SBI and consensus curricula to explore whether evidence exists for differential performances of students based on prior statistical abilities, standardized test score (SAT/ACT) or college GPA.

METHODS

This paper focuses on scores from an adapted version of the CAOS (delMas, Garfield, Ooms, & Chance, 2007) test (see Tintle et al. 2018 for additional details of test construction, administration, reliability and validity on this sample). In particular, our analysis focuses on change in student scores from pre-test (first week of class) to post-test (last week of class) stratifying by levels of student preparation as measured by pretest score, overall GPA or SAT/ACT math score. A primary focus of our analysis is on curricula used by students, comparing simulation-based inference (SBI) and non-simulation-based inference (consensus) curricula. Pre- and post-test data were gathered along with demographic information on students from 43 instructors at 39 institutions using consensus curricula and 83 instructors at 42 institutions using simulation-based curricula. In

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]

addition to exploring overall pre- to post-course change in student scores on the assessment, changes in score will also be looked at within six different subscales of interest: data collection, descriptive statistics, confidence intervals, tests of significance, simulation, and scope of conclusions. Analyses are carried out on 5,266 students who completed both the pre- and post-tests in the 2016/17 academic year. Data cleaning and other data handling processes (including IRB details), are described in our previous work (Chance et al. 2017).

Statistical analyses are performed on changes in test performance pre-course to post-course overall or on subscales of the test. Analyses look at gains pre-test to post-test with matched pairs t-tests. Comparisons between the SBI and consensus curricula are carried out on the pre-test to post-test gains with an independent samples t-test.

RESULTS

Consensus versus SBI on gain as measured by (post-test – pre-test score) within low, middle and high performing students as measured by pre-test score, SAT/ACT z-score and GPA

Students in the SBI curriculum showed similar performance overall on the pre-test compared to students in the consensus curriculum (consensus mean: 45.9% and SBI mean: 47.2%). While overall gains were significant pre-test to post-test within each of the curricula (consensus mean gain: 4.4% and SBI mean gain: 7.2%), the difference in these overall gains was significantly higher for those students using the SBI curriculum (see Table 1). Table 1 also shows these students separated into three groups of approximately equal size (tertiles) based on their pre-test scores.

Table 1: Pre- and post-course test scores stratified by pre-course performance and curriculum

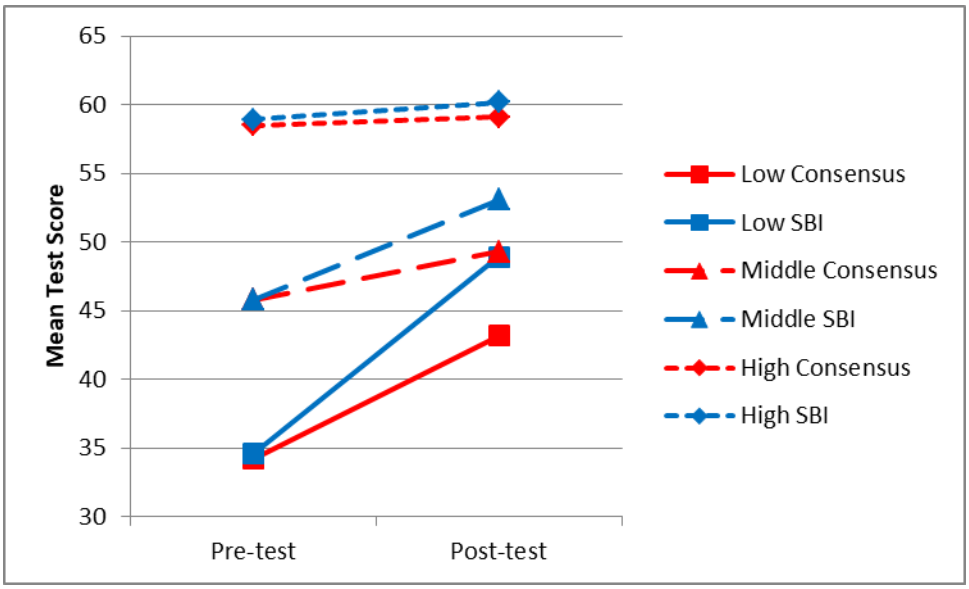
Pre-test score group	**Curriculum**	**Pre-test mean % correct (SD[1])**	**Post-test mean % correct (SD[1])**	**Change in mean % correct (SD[1])[2]**	**Difference in mean change: SBI-consensus (SE)[3]**
Low (<40%)	Consensus (*n*=601)	34.2 (4.8)	43.2 (16.9)	9.0 (16.7)***	5.3 (0.8)***
	SBI(*n*=886)	34.6 (4.6)	48.9 (14.0)	14.3 (14.1)***	
Middle [40% to 50%]	Consensus (*n*=763)	45.8 (3.0)	49.3 (17.1)	3.5 (16.9)***	3.8 (0.7)***
	SBI(n=1353)	45.8 (3.0)	53.1 (14.4)	7.3 (14.3)***	
High (>50%)	Consensus (*n*=562)	58.5 (6.3)	59.1 (17.7)	0.6 (16.6)	0.6 (0.9)
	SBI(*n*=1101)	58.9 (6.4)	60.2 (16.7)	1.3 (16.4)*	
Overall	Consensus (*n*=1926)	45.9 (10.5)	50.3 (18.3)	4.4 (17.1)***	2.8 (0.5)***
	SBI(n=3340)	47.2 (10.5)	54.3 (15.7)	7.2 (15.8)***	

*p<0.05; **p<0.01; ***p<0.001

1. SDs are SDs of student test scores (pre-test or post-test) or SD of change in student test scores.
2. Significance is indicated by asterisks and reported based on results from paired *t*-tests comparing the pre-test and post-test scores
3. SE is from two sample t-test within tertile, significance is indicated by asterisks

As shown in Table 1 and Figure 1, while significant gains are seen within each curriculum (consensus and SBI) for the lowest two tertiles, the highest performers on the pre-test show no significant gains within the consensus curriculum and modest gains within the SBI curriculum. Similarly, it is noted that these gains within the consensus and SBI curricula at the lowest two tertiles differ significantly between the two curricula with students in the SBI curricula outperforming the students in the consensus curricula by between 4 and 5 percentage points. For the highest pre-test performers, there is no statistically significant difference in gains between those using the consensus curriculum and those using the SBI curriculum.

Figure 1. Graph of Pre- and post-course test scores stratified by pre-course performance and curriculum



Tables 2 and 3 mirror Table 1, but stratify students using two different approaches. In Table 2, students are stratified by their self-reported standardized SAT or ACT score. Again, gains within both the consensus and SBI curricula are seen at all three levels except for the lowest level within the consensus curriculum. These gains are once again significantly different between the curricula for the lower two groups with SBI out performing consensus, and no statistically significant difference is found between the curricula in the highest SAT/ACT z-score group, though the SBI curricula students do show larger gains. Table 3 stratifies the data by self-reported college GPA, with similar patterns to Tables 1 and 2.

Table 2: Pre- and post-course test scores stratified by SAT/ACT z-score and curriculum

SAT/ACT z-score Group	**Curriculum**	**Pre-test mean % correct (SD[1])**	**Post-test mean % correct (SD[1])**	**Change in mean % correct (SD[1])[2]**	**Difference in mean change: SBI-consensus (SE)[3]**
Low (<-0.20)	Consensus (*n*=487)	42.1 (9.0)	41.9 (15.6)	-0.2 (16.1)	6.1 (0.8)***
	SBI (*n*=965)	43.5 (9.2)	49.3 (13.5)	5.9 (14.4)***	
Middle [-0.20 to 0.62]	Consensus (*n*=514)	45.8 (10.0)	50.8 (14.4)	5.0 (15.3)***	3.0 (0.8)***
	SBI (n=821)	46.9 (9.3)	54.9 (14.6)	8.0 (15.0)***	
High (>0.62)	Consensus (*n*=618)	50.3 (10.8)	58.5 (18.5)	8.2 (17.8)***	1.6 (0.9)
	SBI (n=850)	51.5 (10.6)	61.2 (15.7)	9.8 (15.2)***	
Overall	Consensus (*n*=1619)	46.4 (10.5)	51.0 (17.8)	4.6 (16.7)***	3.1 (0.5)***
	SBI (*n*=2636)	47.1 (10.3)	54.9 (15.4)	7.8 (14.9)***	

*p<0.05; **p<0.01; ***p<0.001

1. SDs are SDs of student test scores (pre-test or post-test) or SD of change in student test scores.
2. Significance is indicated by asterisks and reported based on results from paired *t*-tests comparing the pre-test and post-test scores
3. SE is from two sample t-test within tertile, significance is indicated by asterisks

Table 3: Pre- and post-course test scores stratified by pre-course self- reported college GPA and curriculum

GPA Group	**Curriculum**	**Pre-test mean % correct (SD[1])**	**Post-test mean % correct (SD[1])**	**Change in mean % correct (SD[1])[2]**	**Difference in mean change: SBI-consensus (SE)[3]**
Low (<3.20)	Consensus (*n*=526)	43.9 (9.2)	42.1 (16.6)	-1.8 (16.0)*	5.7 (0.8)***
	SBI (*n*=814)	45.1 (10.3)	49.0 (13.9)	3.9 (14.5)***	
Middle [3.20 to 3.65]	Consensus (*n*=594)	45.5 (10.1)	49.3 (16.3)	3.8 (16.4)***	3.6 (0.6)***
	SBI (*n*=979)	46.5 (10.1)	53.9 (14.3)	7.4 (14.3)***	
High (>3.65)	Consensus (*n*=550)	48.6 (11.2)	58.6 (18.5)	10.0 (17.0)***	-0.7 (0.9)
	SBI (n=973)	49.7 (10.4)	59.0 (16.2)	9.3 (16.1)***	
Overall	Consensus (*n*=1670)	46.0 (10.4)	50.1 (18.4)	4.1 (17.2)***	3.0 (0.5)***
	SBI (*n*=2766)	47.2 (10.4)	54.3 (15.4)	7.0 (15.2)***	

*p<0.05; **p<0.01; ***p<0.001

1. SDs are SDs of student test scores (pre-test or post-test) or SD of change in student test scores.
2. Significance is indicated by asterisks and reported based on results from paired *t*-tests comparing the pre-test and post-test scores
3. SE is from two sample t-test within tertile, significance is indicated by asterisks

Consensus versus SBI on gain as measured by (post-test – pre-test score) results by subscale of the instrument

Table 4 shows the change in percentage correct as measured by post-test score minus pre-test score for the six subscales of the instrument, stratified by curriculum type for students in the lowest pre-course performance group (less than 40% of the pre-test questions correct). Notably, students using both curricula showed significant improvement on all subscales. However, students in SBI curricula significantly outperformed students in consensus curricula on 5 of the 6 subscales, with the largest differences coming from tests of significance (10.7% larger gain for SBI) and data collection and design (8.9% larger gain for SBI). We do not show details of the same analyses for the middle and high performing students; however, we briefly summarize those results here. The middle pre-course performance group (pre-test score between 40% and 50% correct) showed gains for students using SBI curricula to be significantly greater than gains for students using the consensus curricula in the subscales of tests of significance, data collection and design, and simulation: with 7.5%, 13.0% and 15.8% larger gains, respectively ($p<0.001$ in all three cases). For the highest pre-performance group (more than 50% of pre-test questions correct), data collection and design and tests of significance both showed gains to be significantly greater (9.7% and 3.3%, respectively $p<0.05$) for students in the SBI group compared to those using the consensus curriculum, and the consensus curriculum outperformed the SBI curriculum on descriptive statistics (3.4%, $p<0.05$).

Table 4. Within subscale performance comparing SBI to Consensus curriculum for lower performing students as measured by pre-test percent correct

Subscale (# of items)	**Curriculum**	**Pre-test mean % correct (SD[1])**	**Post-test mean % correct (SD[1])**	**Change in mean % correct (SD[1])[2]**	**Difference in mean change by curriculum (SE)[3]**
Data collection and design	Consensus (*n*=601)	32.4 (24.4)	43.9 (22.7)	11.5 (31.5)***	8.9 (1.8)***
	SBI (*n*=886)	36.4 (25.6)	56.8 (25.1)	20.4 (35.1)***	

(5 items)					
Descriptive statistics (6 items)	Consensus (n=601)	33.3 (16.4)	40.3 (26.4)	7.0 (28.9)***	3.2 (1.4)*
	SBI (*n*=886)	33.9 (17.0)	44.2 (22.6)	10.3 (24.6)***	
Tests of significance (10 items)	Consensus (*n*=601)	40.3 (13.7)	48.8 (23.7)	8.4 (26.2)***	10.7 (1.3)***
	SBI (*n*=886)	39.1 (13.8)	58.3 (21.2)	19.2 (24.5)***	
Confidence Intervals (7 items)	Consensus (*n*=601)	27.2 (15.9)	40.2 (21.8)	13.0 (26.7)***	1.4 (1.4)
	SBI (*n*=886)	27.0 (16.1)	41.4 (20.3)	14.4 (26.4)***	
Scope of Conclusions (2 items)	Consensus (*n*=601)	49.7 (32.9)	58.7 (32.8)	9.0 (44.5)***	6.8 (2.4)**
	SBI (*n*=886)	49.0 (34.2)	64.8 (33.3)	15.7 (45.9)***	
Simulation (5 items)	Consensus (*n*=601)	20.4 (18.6)	29.3 (23.5)	8.9 (36.8)***	4.6 (1.5)**
	SBI (*n*=886)	22.7 (18.8)	36.2 (21.8)	13.5(27.4)***	

*p<0.05; **p<0.01; ***p<0.001

1. SDs are SDs of student test scores (pre-test or post-test) or SD of change in student test scores.
2. Significance is indicated by asterisks and reported based on results from paired *t*-tests comparing the pre-test and post-test scores
3. SE is from two sample t-test within tertile, significance is indicated by asterisks

CONCLUSION

Promising results have been demonstrated in early implementations of SBI curricula with larger gains observed when compared to a consensus curriculum (Tintle, et al., 2012, 2013, 2014). More recent work continues to show promising results in curricular advantages using simulation-based curricula in more diverse samples (Chance, Wong, and Tintle, 2017; Chance & Mcgaughey, 2014). When considering the quantitative maturity of the students entering an introduction to statistics class, one hopes to be able to improve the conceptual understanding of all students; the challenge comes in the form of keeping the highest achievers engaged while not losing those students with less preparation than the rest of their classmates. This was the observed result in an analysis of student performance associations with type of curriculum at two institutions (Tintle et al., 2018).

The results of this data analysis show that regardless of the quantitative maturity of students entering an introduction to statistics class, significant gains can be seen pre-course to post course for low, middle, and high levels of quantitatively mature students as measured by pre-course test score, GPA, and SAT/ACT score. These pre- to post-course gains are seen in both the SBI and consensus curricula. The difference comes in the amount of knowledge gained by those students in using an SBI curriculum. For the lowest two-thirds of quantitatively mature students pre- to post-course gains in content knowledge are significantly higher than their counterparts using a consensus curriculum. For the highest one-third of students there is no statistically significant difference in the overall pre- to post-course gain. When diving deeper into the subscales of the instrument we see a similar story with gains in the SBI curriculum over the consensus curriculum coming mainly from understanding of tests of significance and data collection and design, followed by smaller yet still significant gains in understanding of simulation and scope of conclusions. For the SBI curricula, significant within subscale gains can be seen for all three levels of quantitatively mature students. Higher impact of the SBI curriculum on these subscales is in line with previous research (e.g., Tintle et al. 2011, Tintle et al. 2012, Chance et al. 2017, Tintle et al. 2018) and consistent with the areas of focus of the SBI curriculum.

A few limitations of this analysis are worth mentioning. Self-reported GPA and ACT/SAT score are potentially less reliable than other independent measures and had higher levels of missing data than other variables. On the other hand, stratifying by pre-test score leads to potential regression-to-the-mean effects, especially in the lowest performing group. Notably, similar patterns were seen across all three ways of stratifying the data. Other limitations include the fact that the

pre-test and post-test were typically administered outside of class settings. However, students in both curricula took the test under generally similar conditions. Finally, we note that we have conducted a statistical analysis that ignores potential instructor/institution effects and a host of other demographic variables (e.g., ethnicity; first-generation college student), that could impact conclusions. Future work is exploring the use of more sophisticated statistical methods to control these other sources of variation.

In conclusion, in this large sample of over 5,000 students from numerous institutions we saw that students in simulation-based curricula, especially students with less quantitative maturity, grew more in their performance on a standardized test of statistical thinking than students in consensus curricula. Future work is needed in both longitudinal observational studies and randomized, controlled experiments to better elucidate student learning trajectories, adjust and control for potential instructor and institutional effects, and suggest best-practices for SBI curricula to maximize effectiveness.

ACKNOLWEDGEMENTS

This research was supported by NSF grants NSF/TUES/DUE-Phase II # 1323210.

REFERENCES

Carver R, Everson M, Gabrosek J, Horton N, Lock R, Mocko M, Rossman A, Holmes-Rowell G, Velleman P, Witmer J and Wood B. Guidelines for Assessment and Instruction in Statistics Education: College Report. 2016. American Statistical Association.

Chance, B., & Mcgaughey, K. (2014). Impact of a simulation/randomization-based curriculum on student understanding of p-values and confidence intervals. *Proceedings of the 9th International Conference on Teaching Statistics*, *9*.

Chance, B., Wong, J., & Tintle, N. (2017). Student performance in curricula centered on simulation-based inference: a preliminary report. *Journal of Statistics Education*.

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1), 1–15.

del Mas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(28-58).

Lock R, Lock P, Lock K, Lock E, Lock D. (2017) Statistics: Unlocking the power of data. John Wiley and Sons. Second edition.

Scheaffer, R. (1997). Discussion to new pedagogy and new content: The case of statistics. *International Statistics Review*, *65*(2), 156–8.

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

Tintle, N. L., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2013). Challenging the state of the art in post-introductory statistics: preparation, concepts and pedagogy. *Proceedings of the 59th ISI World Statistics Congress*, (Session IPS032), 295–300.

Tintle N., 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. non-simulation-based inference. https://arxiv.org/abs/1802.09455

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

Tintle, N. L., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., … Vanderstoep, J. (2014). Quantitative evidence for the use of simulation and randomization in the introductory statistics course. *Proceedings of the 9th International Conference on Teaching Statistics*.

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V., & 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., Holmes, V., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, *19*(1), 1–25.