COMPARING STUDENT SUCCESS AND UNDERSTANDING IN INTRODUCTORY STATISTICS UNDER CONSENSUS AND SIMULATION-BASED CURRICULA

LAURA A. HILDRETH
Montana State University
laura.hildreth@montana.edu

JIM ROBISON-COX
Montana State University
jimrc@montana.edu

JADE SCHMIDT
Montana State University
jade.schmidt2@montana.edu

ABSTRACT

This study examines the transferability of results from previous studies of simulation-based curriculum in introductory statistics using data from 3,500 students enrolled in an introductory statistics course at Montana State University from fall 2013 through spring 2016. During this time, four different curricula, a traditional curriculum and three simulation-based curricula, were used. Student success rates and understanding of six key statistical concepts are compared among these curricula using mixed logistic regression. Results indicate that after controlling for salient covariates, differences in student success rates are minimal while student understanding under the simulation-based curricula are similar to or better than student understanding under the traditional curriculum suggesting simulation-based curricula may help increase student understanding of several key statistical concepts.

Keywords: *Statistics education research; Simulation-based inference*

1. INTRODUCTION

For nearly the past twenty years, the introductory statistics course has been taught using a curriculum in which statistical inference is based almost exclusively on the normal distribution (Rossman & Chance, 2014). This "consensus curriculum" is typically comprised of three major units as described by Malone, Gabrosek, Curtiss, and Race (2010): descriptive statistics and study design, probability and sampling distributions, and statistical inference. Under this curriculum, students' focus is primarily on procedures and formulas (Cobb, 2007). In contrast, the GAISE guidelines (Aliaga et al., 2005) emphasize conceptual understanding of statistics and statistical literacy through the use of active learning, real data, and technology in introductory statistics courses to teach statistical concepts and analyze data. Cobb (2007) further challenged the consensus introductory statistics curriculum by suggesting that statistical concepts are best learned through randomization-based inference, now feasible with advances in computing. Following these guidelines, several simulation-based textbooks (e.g., Lock, Lock, Morgan, Lock, & Lock, 2012; Tintle et al., 2016; Zieffler, 2012) have been written. In this paper we refer to these curricula collectively as simulation-based curricula. A number of articles have been written that serve as case studies that describe the development and implementation of simulation-based curricula at various

Statistics Education Research Journal, 17(1), 103-120, http://www.stat.auckland.ac.nz/serj

institutions (cf. Budgett, Pfannkuch, Regan, & Wild, 2013; Fitch & Regan, 2014; Lock, Lock, Morgan, Lock, & Lock, 2014; Roy et al., 2014).

Because these curricular changes are intended to improve student understanding of fundamental statistical concepts, several studies have been conducted to assess the impact of the simulation-based curriculum on student understanding. The work of Tintle and his colleagues provides an in-depth assessment of the effect of simulation-based curriculum on student learning at a single institution. Tintle, VanderStoep, Holmes, Quisenberry, and Swanson (2011) compare pretest and posttest scores on the Comprehensive Assessment of Outcomes in Statistics (CAOS; delMas, Garfield, Ooms, & Chance, 2007) tool for students at Hope College taking a consensus-based introductory course in fall 2007, students at Hope College taking a simulation-based version in fall 2009, and a national sample. Results indicate that CAOS scores increased similarly on all scales measured for both curricula except for statistical inference where gains in student understanding are greater for students under the simulation-based curriculum than those under the consensus curriculum. In a follow-up study of a subset of the students from their 2011 study, Tintle, Topliff, VanderStoep, Holmes, and Swanson (2012) found that students in the simulation-based class had better retention of the material compared to students in the consensus class.

Other previous studies (cf. Chance & McGaughey, 2014; Chance, Wong, & Tintle, 2016; Chance, Holcomb, Rossman, & Cobb, 2010; Holcomb, Chance, Rossman, Tietjen, & Cobb, 2010; Pfannkuch & Budgett, 2014; Taylor & Doehler, 2015; Tintle et al., 2014) have focused on evaluating gains in students' understanding of specific elements of statistical inference such as *p*-values, confidence, and significance. The results of these studies have ranged from neutral to positive in that the gains in students' understanding of statistical inference under simulation-based curricula are comparable to or greater than the gains observed by students taught using the consensus curriculum. Tintle and his colleagues (Swanson, VanderStoep, & Tintle, 2014; Tintle et al., 2014) have also investigated student attitudes towards statistics under the simulation-based curriculum. This work indicates marginal, but practically insignificant, improvements in students' attitudes towards statistics under the simulation-based curriculum in comparison to the consensus curriculum.

Taken altogether, these previous studies show promise for the simulation-based curriculum and bring to light more questions. As noted by Tintle et al. (2014), one such question is: "Are the findings transferable to institutions beyond the single institution described in the initial papers (Tintle et al., 2011, Tintle et al., 2012)?" Concerns exist regarding the generalizability to other institutions due to differences in student populations and instructors across institutions. For example, at larger institutions, introductory statistics courses are typically taught in multiple sections, often by graduate teaching assistants and adjunct faculty who are often not involved in creating the curriculum. This may limit the findings found at larger institutions as the "typical" instructor at these institutions differs considerably from those at smaller institutions.

This study aims to add to and expand the existing literature on the impacts of simulation-based curriculum on student outcomes by using data from Montana State University, a public university of around 16,000 students that offers baccalaureate, master's and doctoral degrees, to assess the impact of simulation-based curriculum on student success and understanding in Stat 216: Introduction to Statistics for students enrolled in the course from fall 2013 through spring 2016. This course is primarily taught by graduate teaching assistants and adjunct faculty who are not primarily responsible for curriculum development. During this time, four different curricula, as described in Section 2, were used to teach this course. Two different student outcomes measures are used. First, we compare student understanding across the four curricula using mixed effects logistic regression. Student understanding was measured using six questions taken directly from or very closely aligned with the CAOS instrument. We then use a mixed effects logistic regression to compare success rates among these four curricula. The choice of response variables and how they and the chosen covariates are measured are explained in Section 2.

2. METHODOLOGY

2.1. PARTICIPANTS FROM STAT 216 AT MSU

The participants in this study were students enrolled in Stat 216: Introduction to Statistics at Montana State University (MSU) from fall semester 2013 through spring semester 2016, excluding summer sessions. During this time, a total of 4,265 students enrolled in this course. As explained below, because our data are from two different sources the sample size for the analysis of student understanding differs from the sample size of the analysis of student success.

2.2. STAT 216 AT MSU

Stat 216 is offered through the Department of Mathematical Sciences at MSU and is the largest course on campus (annual enrollment in Stat 216 is currently around 1,500 students, representing approximately 10% of total student enrollment at MSU). Each semester multiple sections (16 to 22) are taught, each with around 40 students. The course is primarily taught by graduate students and adjunct faculty with sections occasionally taught by tenure track statistics faculty. Classes meet either three times per week for 50 minutes on Mondays, Wednesdays, and Fridays (MWF) or twice a week for 75 minutes on Tuesdays and Thursdays (TR).

Because Stat 216 is required by most degree programs at MSU, it is often seen as a “gatekeeper” course students must pass to proceed in their programs of study. In the past, students have struggled with the course, leading to high withdrawal rates and low grades. Students ended up taking the course multiple times causing concern by the MSU administration about student progress. Under the encouragement of MSU administration we considered different teaching methods and curricula that might improve student outcomes. In particular, we implemented two major changes: choice of curriculum and classroom layout.

Curricula The topics covered in Stat 216, regardless of the curriculum used, include descriptive statistics and statistical inference, including hypothesis tests and confidence intervals for one proportion, one mean, two proportions, two means, and regression slope in simple linear regression. From fall 2013 through spring 2016 these topics were taught using four different curricula as described below.

> ***DVB:*** This consensus curriculum was designed using *Stats: Data and Models*, 3rd edition, by DeVeaux, Velleman, and Bock (2008), covering Chapters 1–22. Class time was primarily devoted to lectures over the material covered in the book. In fall 2013, 13 of the 21 sections were taught using this curriculum and in spring 2014, 7 of the 17 sections used this curriculum.
>
> ***CATALST:*** In spring 2013 we piloted the simulation-based Change Agents for Teaching and Learning Statistics curriculum (CATALST; CATALST, 2012) in a limited number of sections. Starting in fall 2013 this curriculum was more widely implemented with eight out of 21 sections using this curriculum, eight of 17 sections used this curriculum in spring 2014, and in fall 2014 eight out of 22 sections used this curriculum. The CATALST curriculum incorporates model eliciting activities (MEAs) and implements the ideas proposed by Cobb (2007) using a simulation-based approach to understanding inference. This curriculum consists of three units: (1) Chance Models and Simulation, (2) Models for Comparing Groups, and (3) Estimating Models Using Data (Garfield, delMas, & Zieffler, 2012), and uses TinkerPlots^{TM} software (for more details on TinkerPlots, see Konold & Kazak, 2008; Konold et al., 2011; Konold & Miller, 2011) to illustrate statistical concepts via simulation. Because the CATALST curriculum is designed to be used for a terminal statistics course, several modifications were made to the curriculum. In spring 2014, Unit 1 was shortened by replacing the *iPod Shuffle MEA* with an activity entitled *How random is ‘random’ drug testing.* In the original activity, students examine what it means for data to be randomly generated and explored the expected distribution of randomly generated

data. The new activity focuses on similar concepts, but does so in more detail. In Unit 3 the second activity on sampling distributions was used but otherwise this unit was rewritten to formalize the connection between hypothesis tests and confidence intervals by introducing a 95% confidence interval as a list of plausible values that would not be rejected when tested at a 5% significance level. An activity called *Kissing the right way*, which focuses on inference for one proportion, was introduced. Starting in fall 2014 the use of TinkerPlots was replaced with *StatKey* web apps as described below.

Lock: In spring 2014, two sections piloted the use of a simulation-based curriculum based on the Lock et al. (2012) textbook and its associated web apps. This curriculum covered material from Chapters 1–6 and made use of *StatKey*, web apps developed to accompany the textbook. Starting fall 2014, 14 out of 22 sections used this curriculum as it replaced the DVB curriculum. The ordering of topics during fall 2014 and subsequent semesters was changed to be consistent with the ordering of topics used in Tintle et al. (2016). The same curriculum was used in spring 2015 with 9 out of the 17 sections. This curriculum was primarily delivered using lectures with occasional in-class activities.

MSU: Starting spring 2015, we implemented a simulation-based curriculum that we refer to as the MSU curriculum. The topics and order of the topics are the same as under the Lock curriculum with two major changes in how the course is taught. The first is that R-Shiny web apps were developed and used for simulation based-inference (with one app from Rossman and Chance (2008) also being used). These apps, called Sp-IntRo Stats (Robison-Cox, 2016), are found at https://github.com/MTstateIntroStats/SpIntro-Stats. The second major change was in the pedagogy. Classroom time consisted of active learning during which students worked through activities in groups. During spring 2015 eight of the 17 sections used this curriculum. In fall 2015 and spring 2016 all 22 and 18 sections, respectively, used this curriculum. The materials for this course are available as a course pack at https:// github.com/MTstateIntroStats/IntroStatActivities.

Classrooms Courses were taught in one of two types of classrooms—regular or technology enhanced active learning (TEAL) classrooms, which are similar to student-centered active learning environments for undergraduate programs (SCALE-UP) classrooms (for more information see Beichner, 2006) and Beichner and Saul, 2005). Although large computer monitors on the walls make the room look different, the major change in the TEAL classrooms is the use of 7-foot diameter round tables seating nine students that allow students to easily work in three groups of three. Regular classrooms are the traditional lecture style classroom with moveable desks facing the front of the room. From fall 2013 through spring 2015 all sections taught using the CATALST and MSU curricula were taught in TEAL classrooms while the other sections, which were all taught using the DVB and Lock curricula, were taught in regular classrooms. Starting in fall 2015, when all sections used the MSU curriculum, sections taught in the regular classrooms tried to mimic the set-up in TEAL classrooms, if possible, by moving desks into arcs with three students in one group facing three others in another group to form a rough semicircle.

2.3. VARIABLES

Data were obtained from two different sources. Data regarding student understanding were collected by the Stat 216 Student Success Coordinator, a non-tenure track faculty member responsible for coordinating Stat 216 at MSU. These data include measures of student understanding as well as classroom and course characteristics as described below. These data were measured at the section level. Data for the analysis of student success were obtained from the MSU Office of the Registrar. These data include the academic outcome (letter grade or W for withdrew) in Stat 216 and the student characteristics as described below. The data for this analysis are measured at the student level. It is important to note that the data from the Student Success Coordinator and the Office of the Registrar were measured at two different levels. Because of this, in our analysis of student understanding we are only able to control for the classroom and course characteristics described below whereas for the analysis on student outcomes, we are able to control for the student characteristics described below, as well as the classroom and course characteristics. Approval for use of these data in this study was obtained from the Institutional Review Board at MSU.

Student understanding As statisticians we care that our students actually master the concepts of introductory statistics, not just about their final grade. To assess student understanding we use data collected by the Student Success Coordinator. As part of the final exam for each semester, six questions taken either directly from or developed to align with items from the CAOS instrument were asked. Slight modifications were made each semester by changing the scenario presented but not the topic being assessed. The six topics we consider and the item(s) they correspond to are: understanding the purpose of randomization in an experiment (item 7), understanding that correlation does not imply causation (item 22), understanding that lack of "statistical significance" does not guarantee that there is no effect (item 23), ability to recognize correct and incorrect interpretations of *p*-values (item 25, 26, and 27), ability to correctly interpret a confidence interval and to identify misinterpretations of the confidence level (items 28, 29, 30, and 31), and understanding of the factors that allow a sample of data to be generalized to the population of interest (item 38). For each section, the Student Success Coordinator recorded the number of students correctly and incorrectly answering each of the six questions.

Academic outcome in Stat 216 We use the academic outcome in Stat 216 as a measure of student success from which we create a binary variable with two levels: success (final grade of A, B, or C) or non-success (final grade of C- or lower, or withdrew from the course after the 15th day of instruction with approval of the instructor and the student's academic adviser). This variable was chosen as the response as the MSU administration was initially concerned about student success rates in Stat 216. Because assessments given and expectations across semesters and curricula are similar, we feel that comparing success rates is reasonable. Other possible ways of defining academic outcome were also considered but due to the similarity in results and the interests of our administration we ultimately chose to define academic outcome dichotomously as success/non-success.

Classroom and course characteristics Of primary interest in this study is the impact of curriculum on success rates and student understanding. Type of curriculum falls into one of four categories: DVB, CATALST, LOCK, or MSU as explained previously, with MSU serving as the reference group. Type of room the course met in (TEAL or regular) was included using an indicator variable with the regular classroom serving as the reference group. Days of the week the given section met (MWF or TR) was included as an indicator variable with TR serving as the reference group. The time of day the section was taught (morning or afternoon) was included as an indicator variable with morning serving as the reference group. A set of indicator variables was initially considered to model all the different times of day sections meet but ultimately was not chosen as this option requires the use of a relatively large number of indicator variables and does not provide additional information regarding success rates. An indicator variable for term (fall or spring) with spring as the reference group and three indicator variables for year (2013, 2014, 2015, or 2016) with 2016 as the reference group were also included.

It is important to note that there are several potential issues with confounding in this study in regards to classroom level characteristics. The days on which the course was taught and length of course are always confounded. Further, the simulation-based activities were originally developed for use in sections taught on TR and were initially taught exclusively in TEAL classrooms leading to confounding of curriculum, classroom type, and days of the week the class met. Starting fall 2015, all sections, regardless of type of classroom and days of the week the course met, were taught using the MSU curriculum. Sections meeting on TR and two sections meeting on MWF were taught in TEAL classrooms in fall 2015 whereas in spring 2016 more MWF sections, though not all, were taught in TEAL classrooms.

Student characteristics Due to limited demographic information available on the students from the Registrar's Office and the retrospective nature of this study, a limited number of student level characteristics are included in the analysis. Demographic information such as gender and race/ethnicity is not available for this analysis nor do we have baseline measures for student understanding of statistical concepts, such as a pre-assessment given at the beginning of the semester. Though this is somewhat limiting, the student population of MSU is relatively homogeneous and we are able to include several variables to help control for student ability. An exploratory analysis (not included here) of the student characteristics described below indicate that these characteristics are similar over time and across sections.

Class standing, with senior serving as the reference group, and cumulative MSU GPA (on a 4 point scale) , as an overall measure of academic ability, were included. Less than 1% of the students taking Stat 216 were in their first semester and had no MSU GPA. We expect that students in their first semester differ from the other Stat 216 students so we ultimately decided to exclude these students from our analysis on student success.

Mathematical ability is included by using student transcript information: previous math courses taken, grades for each, and scores on standardized exams. Prerequisite math courses, including title and course number, include College Algebra (M121), Math for the Liberal Arts (M145), Language of Mathematics (M147), Secrets of the Infinite (M149), Precalculus (M151), Survey of Calculus (M161), Calculus for Technology I (M165), Calculus for Technology II (M166), and Calculus I (M171) or higher. Nine indicator variables were created for previous math courses taken at MSU where a value of 1 indicates the student had previously taken the given course that was counted as a prerequisite for Stat 216 and earned a grade of C or higher and a 0 indicates that either the student did not take the course or took the course but did not earn a grade of C or higher. [Note, these are not mutually exclusive categories and all nine variables were included in the analysis.]Three more (non-exclusive) indicator variables were created for three standardized tests that students may use to meet the prerequisites for Stat 216 (ACT, SAT, and the Mathematics Placement Exam (MPLEX)—an exam administered at MSU to place students in the appropriate math course). A value of 1 indicates that the student took the exam and met or exceeded the required score to use the standardized exam to satisfy Stat 216 prerequisites (ACT score of 23 or higher, SAT score of 540 or higher, and MPLEX score of 3.5 or higher) whereas a value of 0 indicates that the student either did not take the exam or took the exam and did not earn a score necessary to satisfy the prerequisites for Stat 216.

2.4. STATISTICAL ANALYSIS

We first calculated summary statistics of the student, classroom, and course characteristics using the data provided from the Office of the Registrar (n = 3,491). To assess the effect of curriculum on student understanding, a logistic mixed effects model was run for each of the six assessment items, with classroom and course characteristics as fixed effects, and instructor as a random effect (n = 3,612). To examine the impact of curriculum on success rates, a logistic mixed effects model was run, which included student, classroom, and course characteristics as fixed effects, and instructor as a random effect (n = 3,491). The assumptions for these analyses were evaluated and were reasonably satisfied. All analyses were conducted using SAS 9.4. When reporting our results, we report point estimates and 95% confidence intervals where warranted. We chose to use a multiple comparisons adjustment only when comparing across curricula as this is the focus of this paper. We believe that the emphasis of the analysis should be on the estimated effects (practical significance) of the curricula and not statistical significance. This allows us to meet the overall objective of the paper, which is to compare the curricula and examine how our results compare to previous results.

3. RESULTS

3.1. SUMMARY STATISTICS

Summary statistics of the student, classroom, and course characteristics are provided in Table A1 of the appendix. Means and standard deviations are presented for quantitative variables and counts and percentages are provided for categorical variables. These summary statistics indicate that a majority of Stat 216 students are sophomores, many of whom met the prerequisites for the course by their score on the ACT. The average GPA for the students in our analysis is 2.99 with a standard deviation of 0.72.

3.2. STUDENT UNDERSTANDING

For each section, the number of students answering a given question correctly and incorrectly were reported. Data were not provided for two out of the 116 different sections. In those 114 sections for which data are available, information from 3,612 students is available.

Table 1 displays the percentage of students correctly answering the six questions corresponding to CAOS items for each curriculum and the corresponding chi-square statistic and *p*-value under the null hypothesis that the proportion of students answering the question correctly is the same across all four curricula (all tests are based on 3 df). The number of students taught under each curriculum is provided in parentheses and, for each question, the bold font indicates the curriculum with the highest proportion of students answering the question correctly. Results indicate that for all questions there are differences across the curricula. Interestingly, the curriculum with the highest proportion correct depends on the item, suggesting that there is no one curriculum that is "best" in helping students understand these concepts.

Table 1. Proportion correct for six CAOS items by curriculum

	DVB (n = 84)	CATALST (n = 770)	Lock (n = 758)	MSU (n = 1500)	Total (n = 3612)	χ^2 (df = 3)	*p*-value
Purpose of randomization	37.91	**74.29**	57.12	66.27	61.50	210.62	<0.0001
Association is not causation	**75.59**	69.35	71.24	74.67	72.97	10.52	0.0146
No effect vs. insignificant	73.5	82.18	**88.52**	81.93	82	50.49	<0.0001
Interpretation of *p*-value	70.72	82.08	**83.77**	62.4	72.43	161.14	<0.0001
Interpretation of Conf Intervals	93.49	**94.16**	87.2	92.67	91.97	31.13	<0.0001
Inference to population	74.32	87.01	90.63	**87.67**	86	83.87	<0.0001

We next fit mixed logistic regression models for each item with classroom and course characteristics including curriculum, room type, days of the week the course meets, semester, and year, included as fixed effects, and instructor as a random effect. Because the data were collected at the section level it is not possible to include student characteristics in this analysis. A separate model was fit for each item. The estimated odds ratios and associated 95% confidence intervals associated with the fixed effects in the logistic mixed models are found in Tables 2 and 3. Odds ratios greater than 1 indicate that the odds of correctly answering a question are greater for students in the given group compared to the reference group. For example, the estimated odds ratio associated with the CATALST curriculum for the item on the purpose of randomization is 2.56, which indicates that after controlling for the other variables in the model, the odds of a student taught using the CATALST curriculum answering this question correctly is estimated to be 2.56 times larger than the odds of a student taught under the MSU curriculum answering the question correctly. Odds ratios less than 1 indicate that the odds of correctly answering a question are lower for students in the given level compared to the reference group. For example, the estimated odds ratio associated with the consensus (DVB) curriculum for the item on the purpose of randomization is 0.56 indicating that the odds of a student taught under the DVB curriculum answering the question correctly after controlling for the other variables in the model is estimated to be 0.56 times the odds of a student taught under the MSU curriculum answering the question correctly. Odds ratios whose corresponding confidence intervals do not contain 1 are denoted with an asterisk.

The results from this analysis indicate a relationship between student understanding and classroom and course characteristics that differs by item. For example, the results from understanding the purpose of randomization in an experiment indicate that students taught in TEAL classrooms are less likely to answer that item correctly than students in a traditional classroom (see Table 2) whereas the opposite is true for interpretation of a *p*-value controlling for the other course and classroom characteristics (see Table 3). These results suggest that student understanding of particular statistical concepts depends on more than just the curriculum and that the effect of a given factor depends on the concept of interest. The effect of instructor depends on the concept of interest (respectively, lack of statistical significance does not guarantee that there is no effect: $\chi^2(1) = 1.49$, *p*-value = 0.11; ability to identify correct and incorrect interpretations of a confidence interval: $\chi^2(1) = 0.28$ *p*-value = 0.30; purpose of randomization: $\chi^2(1) = 3.71$, *p*-value = 0.03; identifying correlation does not imply causation: $\chi^2(1) = 3.72$, *p*-value = 0.03; interpreting p-values: $\chi^2(1) = 8.95$, *p*-value < 0.01; identify factors that allow inference to the population: $\chi^2(1) = 9.27$, *p*-value < 0.01).

Table 2. Estimated odds ratios and 95% CIs for first three items (purpose of randomization, correlation does not imply causation, and lack of statistical significance does not guarantee no effect)

	Purpose of Randomization	Association is not causation	No effect vs. Not significant
	Odds Ratio (CI)	Odds Ratio (CI)	Odds Ratio (CI)
Curriculum			
CAT	2.56 (1.58, 4.17)*	0.89 (0.54, 1.47)	0.71 (0.38, 1.35)
DVB	0.56 (0.31, 1.00)	1.66 (0.90, 2.07)	0.55 (0.26, 1.14)
Lock	0.89 (0.59, 1.33)	0.95 (0.63, 1.44)	1.56 (0.95, 2.58)
MSU	1.00 [reference]	1.00 [reference]	1.00 [reference]
Room			
TEAL	0.73 (0.54, 0.99)*	0.91 (0.65, 1.27)	1.20 (0.85, 1.69)
Regular	1.00 [reference]	1.00 [reference]	1.00 [reference]
Year			
2013	0.41 (0.19, 0.88)*	0.29 (0.13, 0.63)*	1.23 (0.48, 3.17)
2014	0.58 (0.33, 1.02)	0.50 (0.28, 0.90)*	1.44 (0.69, 2.99)
2015	0.91 (0.68, 1.22)	0.45 (0.33, 0.62)*	1.01 (0.70, 1.45)
2016	1.00 [reference]	1.00 [reference]	1.00 [reference]
Term			
Fall	0.91 (0.72, 1.15)	1.75 (1.38, 2.21)*	0.96 (0.72, 1.28)
Spring	1.00 [reference]	1.00 [reference]	1.00 [reference]
Day			
MWF	0.80 (0.58, 1.09)	0.85 (0.57, 1.20)	0.93 (0.64, 1.33)
TR	1.00 [reference]	1.00 [reference]	1.00 [reference]

Note. All variables reported were treated as fixed effects and instructor effect was included as a random effect
*denotes odds ratios statistically different than 1 at the 0.05 significance level

Table 3. Estimated odds ratios and 95% CIs for last three items (interpreting p-values, interpreting confidence intervals, and identifying factors that allow inference to the population)

	Interpreting p-values	Interpreting Confidence Intervals	Factors that allow generalization
	Odds Ratio (CI)	Odds Ratio (CI)	Odds Ratio (CI)
Curriculum			
CAT	1.25 (0.69, 2.26)	1.14 (0.50, 2.56)	1.35 (0.63, 2.89)
DVB	0.52 (0.25, 1.05)	1.08 (0.39, 2.94)	0.64 (0.26, 1.56)
Lock	1.78 (1.10, 2.90)*	0.37 (0.19, 0.71)*	1.92 (1.01, 3.63)*
MSU	1.00 [reference]	1.00 [reference]	1.00 [reference]
Room			
TEAL	1.61 (1.17, 2.22)*	0.99 (0.57, 1.73)	1.06 (0.70, 1.59)
Regular	1.00 [reference]	1.00 [reference]	1.00 [reference]
Year			
2013	6.89 (2.82, 16.85)*	0.98 (0.26, 3.67)	0.58 (0.19, 1.79)
2014	4.16 (2.10, 8.23)*	1.77 (0.69, 4.51)	0.58 (0.24, 1.38)
2015	2.27 (1.65, 3.11)*	1.34 (0.80, 2.25)	0.80 (0.52, 1.23)
2016	1.00 [reference]	1.00 [reference]	1.00 [reference]
Term			
Fall	0.53 (0.41, 0.70)*	1.00 (0.65, 1.53)	1.05 (0.75, 1.46)
Spring	1.00 [reference]	1.00 [reference]	1.00 [reference]
Day			
MWF	1.39 (0.98, 1.96)	1.24 (0.70, 2.19)	1.02 (0.65, 1.60)
TR	1.00 [reference]	1.00 [reference]	1.00 [reference]

Note. All variables reported were treated as fixed effects and instructor effect was included as a random effect
*denotes odds ratios statistically different than 1 at the 0.05 significance level

To better compare curricula, we constructed Tukey-Kramer adjusted 95% confidence intervals using the estimates from the mixed logistic regression models for each pair of curricula for each item as displayed in Tables 4 and 5 with significantly different pairs denoted by an asterisk. These results are similar to the results found by Tintle and his colleagues (Tintle et al., 2011; Tintle et al., 2012) in that student understanding under simulation-based curricula is similar to or better than student understanding under a consensus curriculum. These results indicate that there are differences in student understanding among the simulation-based curricula and that the relationship depends on the concept of interest. These results suggest that certain simulation-based curricula are better suited at teaching a given concept and that there is no one “best” curriculum. This is not surprisingly as each curriculum emphasizes different concepts and presents them in different ways which can lead to differences in student understanding. For example, for the question on the purpose of randomization in an experiment a higher proportion of

Table 4. Tukey-Kramer adjusted 95% confidence intervals of odds ratios comparing student understanding by curricula for first three items

	Purpose of Randomization	Association is not causation	Not effect vs. Not Significant
DVB-CAT	(0.14, 0.35)*	(1.12, 3.09)*	(0.46, 1.28)
DVB-Lock	(0.39, 1.01)	(1.04, 2.95)*	(0.19, 0.63)*
DVB-MSU	(0.26, 1.20)	(0.74, 3.71)	(0.21, 1.43)
Lock-CAT	(0.22, 0.54)*	(0.66, 1.73)	(1.23, 3.90)*
Lock-MSU	(0.52, 1.50)	(0.55, 1.64)	(0.81, 3.02)
MSU-CAT	(0.21, 0.74)*	(0.58, 2.16)	(0.61, 3.24)

Note. Estimates were obtained from a model where classroom and course characteristics were fixed effects and instructor was a random effect
*denotes statistically significant different pairs at the 0.05 significance level

Table 5. Tukey-Kramer adjusted 95% confidence intervals of odds ratios comparing student understanding by curricula for last three items

	Interpreting *p*-values	Interpreting Confidence Intervals	Factors that allow generalization
DVB-CAT	(0.24, 0.71)*	(0.40, 2.23)	(0.24, 0.91)*
DVB-Lock	(0.16, 0.52)*	(1.20, 7.00)*	(0.17, 0.66)*
DVB-MSU	(0.20, 1.32)	(0.29, 4.02)	(0.20, 2.05)
Lock-CAT	(0.84, 2.42)	(0.05, 0.73)*	(0.73, 2.77)
Lock-MSU	(0.94, 3.38)	(0.16, 0.87)*	(0.83, 4.43)
MSU-CAT	(0.37, 1.75)	(0.30, 2.56)	(0.27, 2.01)

Note. Estimates were obtained from a model where classroom and course characteristics were fixed effects and instructor was a random effect

*denotes statistically significant different pairs at the 0.05 significance level

students taught with the CATALST curriculum answered the question correctly than students taught using any of the other three curricula and the proportions of students answering this question correctly in the non-CATALST curricula were similar. For the question on recognizing that lack of statistical significance does not guarantee no effect a higher proportion of students taught under the Lock and MSU curricula answered the question correctly than students taught under the CATALST or consensus curricula.

3.3. STUDENT SUCCESS

For the analysis of student success, we used student-level data from the Office of the Registrar. We included undergraduate students attempting the course for the first time and not receiving an incomplete (I) grade for a sample size of 3,491 students. It is important to note that this sample of students is not the same as the sample of students used in the analysis of student understanding. In the analysis of student success we have student-level information available allowing us to exclude students repeating the course while including students who withdrew from the course. Consequently, the same students are not included in both analyses so comparing results from the two analyses must be done cautiously.

Of interest in this analysis is the relationship between curriculum and success rates, as displayed in Table 6. For each curriculum, the number and percentage of students successful and not successful in the course are reported. Most noticeable is that the success rate is higher for both the MSU and the CATALST curricula, which both have success rates over 81%, compared to the DVB and Lock curricula which have success rates below 75%. A chi-square test of independence ($\chi^2(3) = 52.584$, *p*-value < 0.0001) suggests that there is a relationship between success rates and curriculum, though it does not account for other potential causes and thus should be interpreted with caution.

Table 6. Success rates by curriculum

Curriculum	Non-Success *n* (%)	Success *n* (%)
CATALST	98 (14.0)	601 (86.0)
DVB	158 (28.6)	395 (71.4)
Lock	193 (25.1)	575 (74.9)
MSU	275 (18.7)	1196 (81.3)
Total	724 (20.7)	2767 (79.3)

We then fit a mixed logistic regression model with student characteristics (GPA, class standing, math course history, standardized exam history) and classroom and course characteristics (curriculum, room type, time and days of the week the course meets, semester, and year) included as fixed effects and instructor is treated as a random effect. Estimated odds ratios and the associated 95% confidence intervals of the fixed effects for the logistic mixed effects model are presented in Table 7 with odds ratios significantly different than 1 at the 5% significance level denoted by an asterisk. Estimated odds ratios greater than 1 indicate that the odds of success are greater for students in the given group than compared to the reference group while the opposite is true for estimated odds ratios of less than 1. Of note is that GPA, not surprisingly, has a considerable impact on student success rates with an estimated odds ratio of 13.111 with a 95% confidence interval of 10.551 to 16.290. Class standing is also related to success rates with subsequent analyses (not shown) indicating that juniors are more likely to be successful in the course

Table 7. Estimated odds ratios and 95% CIs of the fixed effects for student success

Variable	Odds Ratio (CI)
GPA	13.11(10.55, 16.29)*
Class	
Freshman	0.65 (0.40, 1.06)
Sophomore	0.67 (0.45, 1.00)
Junior	1.18 (0.77, 1.82)
Senior	1.00 [reference]
Math History Prerequisites	
College Algebra (M121)	0.74 (0.57, 0.95)*
Math for the Liberal Arts (M145)	0.58 (0.38, 0.88)*
Language of Mathematics (M147)	1.50 (0.95, 2.35)
Secrets of the Infinite (M149)	0.62 (0.23, 1.67)
Precalculus (M151)	0.05 (0.00, 0.88)*
Survey of Calculus (M161)	1.21 (0.92, 1.59)
Calculus for Technology I (M165)	2.35 (0.58, 9.46)
Calculus for Technology II M166	2.26 (0.57, 8.96)
Calculus 1 (M171) or higher	1.17 (0.78, 1.75)
Standardized Test Prerequisites	
SAT	1.00 (0.76, 1.33)
ACT	1.56 (1.23, 1.99)*
MPLEX	0.97 (0.67, 1.41)
Curriculum	
CAT	2.70 (0.92, 4.64)
DVB	1.52 (0.55, 4.25)
Lock	0.72 (0.38, 1.37)
MSU	1.00 [reference]
Room	
TEAL	1.73 (0.96, 3.10)
Regular	1.00 [reference]
Days of the Week	
MWF	1.60 (0.89, 2.87)
TR	1.00 [reference]
Year	
2013	0.46 (0.13, 1.60)
2014	0.87 (0.35, 2.14)
2015	0.83 (0.51, 1.34)
2016	1.00 [reference]
Term	
Fall	1.44 (0.96, 2.17)
Spring	1.00 [reference]
Time	
Afternoon	0.74 (0.57, 0.98)*
Morning	1.00 [reference]

Note. All variables reported were treated as fixed effects and instructor effect was included as a random effect
*denotes an odds ratio statistically different than 1 at the 0.05 significance level

than sophomores. There was also a strong instructor effect ($\chi^2(1)= 36.96$, p-value < 0.0001) which is not shown in Table 6.

The indicators of student prerequisite checks for math history, and standardized tests reveal several unexpected results. Notably, the odds ratios for the indicators for College Algebra (M121), Math for the Liberal Arts (M145), and Precalculus (M151) are less than 1 with associated confidence intervals not containing 1 indicating that students that have completed these courses successfully are less likely to successfully complete Stat 216 compared to students that either did not take the course or did not successfully complete it. These results must be viewed as conditional on the other variables in the model including GPA and standardized test prerequisite checks, which all are measuring similar concepts and may explain the unexpected results. Further, students who used ACT to satisfy the prerequisite for the course were more likely to be successful in part because they entered college with strong enough math skills to test out of the 100 level math prerequisites, which may contribute to the unexpected estimated odds ratios. Several of the classroom characteristics including type of room, term, year, and days of the week are not useful in predicting success (given the other terms in the model) though we see a small effect for time of day the course meets.

To further explore the impact of curriculum on success rates, 95% Tukey-Kramer adjusted confidence intervals of success rates between curricula calculated from the estimated model were constructed and are displayed in Table 8 with statistically significant different pairs denoted with an asterisk. Only one confidence interval does not contain 1—the comparison of Lock and CATALST which indicates that success rates are lower for students under the Lock curriculum than the CATALST curriculum. Though these results are not as dramatic as those shown in Table 6, the results still suggest that the differences in success rates across curricula are small and that the use of simulation-based curriculum is not detrimental to students' success in Stat 216.

Table 8. Tukey-Kramer adjusted 95% confidence intervals of odds ratios comparing success rates by curricula

Curricula	Confidence Interval
DVB-CAT	(0.29, 1.84)
DVB-Lock	(0.89, 5.03)
DVB-MSU	(0.40, 5.47)
Lock-CAT	(0.16, 0.76)*
Lock-MSU	(0.31, 1.67)
MSU-CAT	(0.17, 1.39)

4. DISCUSSION

This study reports the results of an analysis comparing student success rates and student understanding across four different curricula that were implemented from fall 2013 through spring 2016 at Montana State University. During this time, we used a consensus curriculum (De Veaux, Velleman, & Bock, 2008), the CATALST curriculum, a curriculum based on Lock et al. (2012), and the MSU curriculum. When comparing student understanding by curriculum there are pronounced differences. When not controlling for other covariates, we note that there are considerable differences in student understanding by curriculum and that no one curriculum outperforms the others. When including classroom and course characteristics using a logistic mixed effects model, these differences become slightly less pronounced. We notice that for all items, student understanding under a simulation-based curriculum is similar to or better than student understanding under the consensus curriculum. How the three simulation-based curricula compare depends on the statistical concept. Specifically, for the question on the purpose of randomization, students taught under the CATALST curriculum were more likely to correctly answer this question than students taught with any of the other three curricula. For the question on identifying that correlation does not imply causation students taught using the consensus (DVB) curriculum were more likely to correctly answer this question than students taught using the CATALST or Lock curricula. Conversely, students taught using the consensus (DVB) curriculum were less likely to answer the questions on interpreting *p*-values and identifying factors that allow for inference to the population than students from the CATALST and Lock curricula. These results also indicate that students taught using the Lock curriculum were more likely to correctly answer the question on recognizing that lack of statistical significance does not guarantee no effect than students in the CATALST and consensus (DVB) curricula whereas for the question on interpreting confidence intervals students using the Lock curriculum were less likely to correctly answer this question than students taught under the other three curricula.

When comparing success rates between curricula, student success rates under the CATALST and MSU curricula are nearly 10 percentage points higher than the success rates for students under the consensus and Lock curricula. When including student level and classroom level covariates in the analysis via a logistic mixed effects model, the effect of curriculum is greatly reduced. Based on Tukey-Kramer adjusted 95% confidence intervals, only one pair of curricula—CATALST versus Lock—had a confidence interval of odds ratios which did not contain 1. The results suggest that the probability of successfully completing the course is higher under the CATALST curriculum than the Lock curriculum though differences in student success rates by curriculum are small.

The results of this study add to and broaden the existing body of literature on simulation-based curricula in introductory statistics courses. One question raised by Tintle (2014) was the transferability of the findings in previous studies to different institutions. This study was conducted at a university of around 16,000 students and involved all introductory statistics instructors at the institution allowing for a better understanding of the effectiveness of simulation-based curricula at an institution that differs considerably from the institution in the seminal papers. Our results are consistent with the findings from those initial papers (Tintle et al., 2011; Tintle et al., 2012) in that under a simulation-based curriculum, student understanding is similar to or better than what is observed under a consensus curriculum. Our results are also able to provide further insights into differences among different simulation-based curricula. This study indicates that differences in student understanding among the three simulation-based curricula depends on the concept of interest.

4.1. LIMITATIONS

The findings of this study need to be interpreted thoughtfully as there are several limitations to the study. One potential limitation is that the data available to us was somewhat limited. In the analysis of student understanding, we were unable to include student information because the data were collected at the section level, not the student level. Additionally, for this analysis we only evaluated six concepts. Though we feel that these six concepts are fundamental concepts in introductory statistics, there are other concepts that may be of interest that were not evaluated. In the analysis of student success we had limited student information available and potentially important student level characteristics such as gender, socioeconomic status, and attitudes towards statistics are not available. For both analyses, instructor characteristics, such as teaching experience, were not available to us. Despite the limitations of the data, this analysis does further the understanding of the effectiveness of simulation-based curricula and provides a comparison of different simulation-based curricula on student understanding of several statistical concepts.

For both analyses, there are issues with confounding that we were unable to avoid when implementing the curricula. From fall 2013 through spring 2015, sections taught using the consensus curriculum were taught on MWF in traditional classrooms while sections taught using simulation-based curricula were taught on TR in TEAL classrooms making it difficult to disentangle the separate effects of classroom, curriculum, and days of the week the course was taught. Further, the CATALST and MSU curricula are activity-based while the Lock and consensus curricula were taught primarily through lecture. It is important to take this into account, as differences among curricula may be associated with whether or not active learning was implemented. Lastly, under the MSU curriculum we implemented the use of a ‘helper,’ an upper-level undergraduate or first-year graduate student in statistics, who attended class and interacted with each group. The presence of this helper cannot be controlled for in the analysis as the helper usually was assigned to one section which leads to confounding of the instructor and helper and the presence of a helper is also confounded with the MSU curriculum.

One concern we have when comparing the curricula is using the CATALST curriculum. This curriculum was designed to be a terminal course in statistics whereas the other three curricula are not. This leads to different learning goals and consequently the use of different assessments for the CATALST curriculum. This makes comparisons to the CATALST curriculum questionable as it is not an ‘apples to apples’ comparison. We believe that comparisons among the other three curricula are reasonable to make as the same (or very similar) assessments were used and were graded in the same manner by an overlapping group of instructors. When taking this into consideration we can conclude from the results that there is no evidence that the newer curricula have reduced student success rates.

4.2. IMPLICATIONS FOR RESEARCH

Based on the findings and limitations in this study, there are several potential implications for research. Because the initial studies of Tintle et al. (2011, 2012) were conducted at a small liberal arts college, questions have been raised about the generalizability of the results in these papers in that students taught under simulation-based curricula perform similar to or better than students taught using the consensus curriculum. The results of our analysis are consistent with the results found in Tintle et al. and in Chance, Wong, & Tintle (2017), providing some evidence of the generalizability of these results. To further provide evidence of the generalizability of these findings, it is crucial for other institutions to conduct similar studies if possible. This would allow for greater information regarding the suitability of using simulation-based methods for different student populations and types of institutions. The limitations in this study provide insight regarding how to conduct the study and what data to collect in future studies. Having experienced confounding of curriculum with classroom type, days of the week, and the use of active learning, we recommend that others avoid this so that better estimates of the impacts of simulation-based curricula can be obtained. We also recommend that, if possible, more data be collected from the students including demographic information such as gender, measures of student ability, and measures of student attitudes towards statistics both before and after the course. Due to the retrospective nature of this study we were unable to collect this information from students. The inclusion of these variables is of interest and would provide useful information to statistics educators. Lastly, we also recommend that careful consideration be made about how student understanding is evaluated. The six concepts we evaluated are fundamental concepts in introductory statistics but are not the only concepts taught in the course and are not the only concepts measured by the CAOS instrument. The inclusion of more questions to assess additional concepts would provide further insight on the efficacy of simulation-based curricula.

4.3. IMPLICATIONS FOR TEACHING

With the implementation of a simulation-based curriculum there are three major aspects of the course we have changed: the topics chosen and the ordering of topics, the use of web apps, and the use of group activities. Individual instructors may have different views of the importance of these three aspects, but our collective opinion is that these changes have been beneficial to students. The use of the technology allows students to interact with the concepts thereby enhancing their understanding. We also have reordered the topics dramatically compared to the ordering of topics in the consensus curriculum. The use of randomization allows for the introduction of inference in the first week of class rather than waiting until the second half of the course as we did with the consensus curriculum. We are then able to reinforce the ideas of *p*-value and confidence interval interpretations throughout the course, which helps to improve student understanding of these key statistical concepts.

Based on our experiences, we note that it is important that the instructor carefully monitor the students during group activities as it is not uncommon for a group to come to the wrong conclusion. We have found it beneficial to have a second observer in each class to interact with each group and to ensure that all learn the lesson as planned. We been able to hire undergraduate teaching assistants who act as assistant instructors and allow greater interaction between teachers and students.

We encourage others to try the innovative approaches introduced by simulation-based curricula as we have noticed an enriched experience for students. We have noticed that discussions can reach a greater depth due to the use of group activities leading to deeper understanding. One area that we have noticed improved understanding is limitations of statistical inference and how study design can inhibit the scope of inference. We have also noticed a major change due to the use of groups is that students are held individually accountable. We do see improvements in attendance when students are assigned to groups and when they must work with their group each class period. Overall we have found the use of simulation-based curricula highly beneficial to our students and will continue to use and refine this curriculum in order to improve student learning.

ACKNOWLEDGEMENTS

Many people have contributed to the work of curricular change at MSU. At the risk of missing some people, we would like to acknowledge Megan Higgs who started the quest for a more effective curriculum, Heidi Lindsay who implemented the CATALST curriculum, MSU administration (especially David Singel) who found resources to fund the changes, the many instructors of STAT 216 who tested the different curricula, and the statistics group at MSU for allowing experimentation with this critical course. We would also like to thank Jennifer Green for her insightful comments on this paper and two anonymous reviewers and the editors for their helpful feedback.

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., … Witmer, J. (2005). *Guidelines for assessment and instruction in statistics education: College report.* Alexandria, VA: American Statistical Association.
[Online: http://www.amstat.org/education/gaise/GAISECollege.htm]

Beichner, R. (2006). North Carolina State University: SCALE-UP. In D. Oblinger J.K. Lippincott (Eds.). *Learning Spaces* (29.1–29.6). Boulder, CO: Educause.

Beichner, R. J., & Saul, J.M. (2005). Introduction to the SCALE-UP (Student-Centered activities for large enrollment undergraduate programs) project. In *Invention and impact: Building excellence in undergraduate Science, Technology, Engineering and Mathematics (STEM) education* (pp. 61–66). Washington, DC: American Association for the Advancement of Science.

Budgett, S., Pfannkuch, M., Regan, M., & Wild, C. J. (2013). Dynamic visualizations and the randomization test. *Technology Innovations in Statistics Education*, *7*(2), 1–21.
[Online: https://escholarship.org/uc/item/9dg6h7wb]

CATALST (2012). Change Agents for Teaching and Learning Statistics (CATALST) materials.
[Online: https://github.com/zief0002/Statistical-Thinking]

Chance, B., Holcomb, J., Rossman, A., & Cobb, G. (2010). Assessing student learning about statistical inference. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8)*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://iase-web.org/documents/papers/icots8/ICOTS8_5F1_CHANCE.pdf]

Chance, B., & McGaughey, K. (2014). Impact of a simulation/simulation-based curriculum on student understanding of *p*-values and confidence intervals. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9),* Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://icots.info/9/proceedings/pdfs/ICOTS9_6B1_CHANCE.pdf]

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

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum. *Technology Innovations in Statistics Education*, *1*, 1–15.
[Online: https://escholarship.org/uc/item/6hb3k0nz]

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.
[Online: https://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf]

de Veaux, R. D., Velleman, P. F., & Bock, D. E. (2008). *Intro Stats* (3rd Ed.). Boston, MA: Pearson.

Fitch, A. M., & Regan, M. (2014). Accepting the challenge: Constructing a randomization pathway for inference into our traditional introductory course. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9),* Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://icots.info/9/proceedings/pdfs/ICOTS9_4A1_FITCH.pdf]

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM Mathematics Education*, *44*(7): 883–898.

Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomization tests. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8)*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://icots.info/8/cd/pdfs/invited/ICOTS8_8D1_HOLCOMB.pdf]

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education, 2*(1).
[Online: https://escholarship.org/uc/item/38p7c94v]

Konold, C., Madden, S., Pollatsek, A., Pfannkuch, M., Wild, C., Ziedins, I., … Kazak, S. (2011). Conceptual challenges in coordinating theoretical and data-centered estimates of probability. *Mathematical Thinking and Learning, 13*(1-2), 68–86.

Konold, C., & Miller, C. (2011). *TinkerPlots: Dynamic data exploration* (Version 2) [Computer software]. Emeryville, CA: Key Curriculum Press.

Lock, R., Lock, P., Morgan, K. L., Lock, E. F., & Lock, D. F. (2012). *Unlocking the Power of Data*. Hoboken, NJ: Wiley.

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

Malone, C., Gabrosek, J., Curtiss, P., & Race, M. (2010). Resequencing topics in an introductory applied statistics course. *The American Statistician, 64*(1), 52–58.

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

Robison-Cox, J. (2016). SpIntro-Stats. [Computer application]. Montana State University.
[Online: https://github.com/MTstateIntroStats/SpIntro-Stats]

Robison-Cox, J. (2016). *Stat 216 course pack fall 2016*, Montana State University.
[Online: github.com/MTstateIntroStats/IntroStatActivities/blob/master/coursePack/coursePack.pdf]

Rossman, A., & Chance, B. (2008). *Concepts of statistical inference: A simulation-based curriculum*. [Online: http://statweb.calpoly.edu/csi]

Rossman, A., & Chance, B. (2014). Using simulation-based inference for learning introductory statistics. *WIREs Computation Statistics*, *6*(4), 211–221.

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

Swanson, T., VanderStoep, J., & Tintle, N. (2014). Student attitudes toward statistics from a simulation-based curriculum. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://icots.info/9/proceedings/pdfs/ICOTS9_1F1_SWANSON.pdf]

Taylor, L., & Doehler, K. (2015). Reinforcing sampling distributions through a simulation-based activity for introducing ANOVA. *Journal of Statistics Education*, *23*(3).
[Online: http://ww2.amstat.org/publications/jse/v23n3/taylor.pdf]

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VandersStoep, J. (2016). *Introduction to Statistical Investigations*. Hoboken, NJ: Wiley.

Tintle, N., 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. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8A3_TINTLE.pdf]

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary simulation-based introductory statistics curriculum. *Statistics Education Research Journal*, *11*(1): 21–40.
[Online: https://iase-web.org/documents/SERJ/SERJ11(1)_Tintle.pdf]

Tintle, N., VanderStoep, J., Holmes, V.-L, Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary simulation-based introductory statistics curriculum. *Journal of Statistics Education*, *19*(1).
[Online: http://ww2.amstat.org/publications/jse/v19n1/tintle.pdf]

Zieffer, A. (2012). *Statistical Thinking: A Simulation Approach to Modeling Uncertainty*. Minneapolis, MN: Catalyst Press.

LAURA A HILDRETH
Montana State University
Department of Mathematical Sciences
PO BOX 172400
Bozeman, MT 59717

APPENDIX A: SUMMARY STATISTICS FOR STUDENT, CLASSROOM, AND COURSE CHARACTERISTICS

Table A1: Summary statistics for student, classroom, and course characteristics

Variable	*n*	%
Class Standing		
Freshman	583	16.70
Sophomore	1676	48.01
Junior	891	25.52
Senior	341	9.77
Standardized Test Prerequisites[a]		
SAT	823	23.57
ACT	1871	53.59
MPLEX	341	9.77
Math History Prerequisites[a]		
College Algebra (M121)	964	27.61
Math for the Liberal Arts (M145)	192	5.50
Language of Mathematics (M147)	283	8.11
Secrets of the Infinite (M149)	35	1.00
Precalculus (M151)	2	0.06
Survey of Calculus (M161)	753	21.57
Calculus for Technology I (M165)	63	1.80
Calculus for Technology II M166	70	2.01
Calculus I (M171) or higher	464	13.29
Type of Room		
TEAL	1613	46.20
Regular	1878	53.80
Curriculum		
CATALST	699	20.02
DVB	553	15.84
LOCK	768	22.00
MSU	1471	42.14
Time of Day		
Morning	1706	48.87
Afternoon	1785	51.13
Days of the Week		
MWF	2012	57.63
TR	1479	42.37
Year		
2013	589	16.87
2014	1165	33.37
2015	1195	34.23
2016	542	15.53
Term		
Fall	2003	57.38
Spring	1488	42.62
GPA	*M* = 2.992	*SD* = 0.715

[a] Percentages for these two variables do not sum to 100% as students can take no or multiple standardized tests or have taken none or multiple prerequisite courses.