INFUSION OF DATA SCIENCE AND COMPUTATION IN INTRODUCTORY STATISTICS

Sayed A Mostafa, Tamer M Elbayoumi and Seongtae Kim
North Carolina A&T State University, USA
sabdelmegeed@ncat.edu

This study investigates the impact of infusing data science (DS) knowledge and computational tools in introductory statistics on students' statistical gains and levels of DS awareness, aspirations, and readiness. The data were collected from introductory statistics sections, that used one of two comparative course designs, at a minority-serving university in the United States. In the traditional course sections, students took lecture-style classes and used a calculator for all course computations. In the DS-infused course sections, in addition to the regular class sessions, students engaged in a weekly virtual statistical computing lab via posit Cloud, pre-lab interactive R shiny reading assignments, and a data analysis project with R Markdown. While the results of pre/post content tests did not show positive impacts for the DS-infused design on students' learning gains, the survey results showed positive impacts for the DS-infused design on students' levels of DS awareness and readiness.

INTRODUCTION

Since the publication of Nolan and Temple Lang's (2010) paper on "Computing in the Statistics Curriculum", many statisticians started advocating integrating computing into statistics courses starting with the Introductory Statistics (Intro Stats) course. This trend was further supported by the fast-growing demands on graduates with computational and data analytical skills who can work as data scientists. This has led to a significant body of literature on integrating computing in statistics courses (see Horton and Hardin, 2021, and references therein). With the Intro Stats course being the main, and sometimes the only, source of quantitative training for most undergraduates in the United States, special efforts have been focused on enhancing Intro Stats courses to better prepare students for using solid statistical reasoning in their majors and career fields and to help promote data literacy. Horton et al. (2015) emphasized the importance of computational and data-analytical skills being developed at the early stages of college education and reaching a broad spectrum of students. This study aims to introduce and evaluate an Intro Stats course design that integrates computing as a core component of the course for enhancing the students' statistical gains and boosting their levels of data science (DS) awareness, aspiration, and readiness.

STUDY SETTING AND PROPOSED COURSE DESIGN

The Intro Stats course considered in this study is an algebra-based "Introduction to Probability & Statistics" course that serves students from a wide range of STEM (~46%) and non-STEM (~54%) fields at a large Historically Black University (HBCU) in the United States. Most students in the course are from groups that are known to be underrepresented in Statistics and DS (~82% are African Americans and ~69% are females). About 7 sections of the course are offered every Fall and Spring semester with nearly 30 to 40 students in each section. In the Summer of 2021, we were awarded a National Science Foundation grant to redesign the course to enhance students' statistical gains from the course and attract and prepare our students to pursue data science education and careers (see more project details at https://introtostatncat.github.io). The redesigned Intro Stats course, which we name *DS-infused Intro Stats*, was developed in Fall 2021 and piloted in Spring 2022. Below we describe the traditional and the DS-infused Intro Stats designs under study.

Traditional Intro Stats Course

Traditionally, the Intro Stats course at the study institution consists of 3 hours of lectures weekly. The course is taught in the "traditional manner" where the instructor uses whiteboard or PowerPoint slides or a combination to present definitions, formulas, and procedural steps, and guide students to solve practice problems. All course computations are performed with the assistance of a calculator and/or statistical distribution tables. This course design is consistent with what is known in the literature as the "consensus" Intro Stats course (e.g., Cobb, 2015). The course adopted *OpenIntro Statistics* as the textbook and covered the following topics: data basics; data collections methods; summarizing numerical and categorical data; intro to probability; distributions of random variables and sampling distributions; one and two-sample inference for means and proportions; and intro to linear regression. To counterbalance the weekly time commitment of the traditional and DS-infused sections, an optional 1-hour recitation session is made available to students in the traditional sections.

DS-Infused Intro Stats Course

The DS-infused Intro Stats course has 3 hours of class sessions weekly and adopted the same textbook and covered the same topics as the traditional course described above. The infusion of data science knowledge and computational tools in the Intro Stats course design is displayed in Fig. 1. A major addition to the course was in the form of a 1-hour virtual statistical computing lab that meets weekly via Zoom and uses the posit cloud. During lab sessions, students are guided to complete R coding tasks practicing the concepts and methods covered in class sessions. Lab sessions are facilitated by the section instructor and a graduate teaching assistant. Before lab sessions, students complete assigned interactive R shiny tutorials involving reviewing concepts from class and running R codes with immediate feedback through the *learnr* package tools. At the end of each lab session, students submit a pdf lab report written using R Markdown. Additionally, the computing lab guides students to complete a semester-long major-related data analysis project. This course design aims to expose students, early and frequently, to the elements of the DS workflow to set the stage for further DS education as suggested by Horton et al. (2015) and the authors of the *Journal of Statistics and Data Science Education*'s special issue on Integrating Computing in the Statistics and Data Science Curriculum (Horton and Hardin, 2021). Furthermore, this design fosters active learning in the lab and allows us to help students explore concepts and analyze real data using technology as recommended by the Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report endorsed by the American Statistical Association (GAISE, 2016). Finally, the DS-infused course entices students to pursue Statistics/DS education and careers by engaging them in online discussion board forums promoting the power of statistics and DS for solving real-world problems and exposing them to blog posts about DS educational opportunities and current trends in the DS job market.

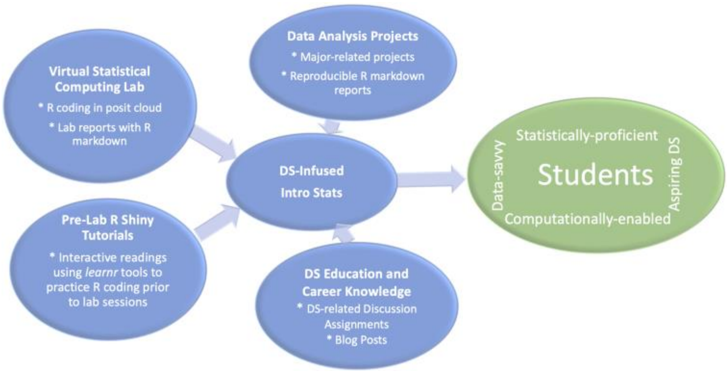

Fig. 1: An Intro Stats course design integrating data science knowledge and computing in the course.

DATA AND METHODS

We used data from both pre/post-concept inventory tests and pre/post-surveys to evaluate the efficacy of the DS-Infused Intro Stats course design in enhancing students' statistical gains from the course and raising their levels of data science awareness, aspiration, and readiness. Data collection was approved by the university's IRB #19-0108.

Students' Performance Data

To assess students' learning gains from the Intro Stats course, we used the revised Comprehensive Assessment of Outcomes in Statistics (CAOS) test. CAOS was originally developed and validated by delMas et al., (2007) and was revised by Tintle et al., (2018). The revised CAOS test consists of 33 questions assessing concepts commonly covered in the Intro Stats course. The instrument was built as a test in the LMS and the students in 4 Intro Stats sections (2 sections taught with the traditional course design and 2 sections with the DS-infused design) completed the test both at the beginning and at the end of the Fall 2022 semester. Both the pretest and posttest were completed during class (~40 minutes). Participation points were given as an incentive for completion.

Data Science Awareness, Aspirations, and Readiness Survey

We developed a survey to assess Intro Stats students' data science awareness, aspirations, and readiness. The survey questionnaire included 3 items about students' awareness of data science educational opportunities at the study institution and 4 items about their awareness of general data scientists' skills and careers, 4 items gauging students' aspirations of data science education, and five items on students' readiness for further statistics/data science education. See the items listed in Tables 2 and 3 below. The questionnaire was piloted and revised in Spring 2022, and it was then used for data collection in 4 sections of Intro Stats in Fall 2022. The survey is administered via Qualtrics: pre-survey during the first week of the semester and post-survey during the second-to-last week of the semester. Students were encouraged to complete the survey outside of class for participation points.

Statistical Analysis

All data were processed and analyzed using the R software (R Foundation for Statistical Computing). Factor analysis of survey items was conducted using the *psych* package. Linear regressions were used to assess the associations between the change in scores (post-score – pre-score) as the response variable and course type (DS-Infused vs Traditional), demographic covariates, and academic factors as explanatory variables. A 5% significance level was used throughout the paper.

RESULTS

Impact on Students' Statistical Gains

The results from the CAOS pre/posttest are presented in Table 1 for Fall 2022. The pretest mean, posttest mean, and average gain (posttest – pretest) are reported for each of the two designs and grouped by several demographic and academic factors. Overall, we observe statistically significant, but modest learning gains regardless of the course design (5.32 vs 5.2 for DS-infused vs traditional, respectively). One possible reason for the modest gains is the low posttest scores possibly due to students choosing to guess on the pretest and posttest as they receive participation points for test completion, not performance. Similar statistically significant learning gains are observed for certain student groups under each of the two designs with the most notable gains observed for non-STEM students in both course types, and non-financial aid and high-achieving students in the traditional sections. However, no significant associations were observed in the multiple regression model for the gain on course type and students' demographics/academic factors. These results should be interpreted with caution as the sample size is relatively low to draw a decisive conclusion about the efficacy of the DS-infused design in boosting students' statistical gains. Furthermore, the DS-infused course is mainly designed to enhance students' computational competencies which are not measured by CAOS.

Table 1. CAOS test results in two versions of the Intro Stats course.

		DS-Infused				Traditional[§]				
		n	Pretest mean (SD)	Posttest mean (SD)	Mean diff (SD)	n	Pretest mean (SD)	Posttest mean (SD)	Mean diff (SD)	Coeff. (SD)[§§]
Course Type[‡]		57	39.5 (11.2)	44.8 (12.2)	5.3 (12.9)**	49	42.4 (11.1)	47.6 (12.6)	5.2 (11.8)**	0.9 (2.8)
Sex	Female[§]	38	38.0 (11.6)	43.5 (11.8)	5.5 (12.6)**	32	40.5 (9.6)	46.8 (13.5)	6.3 (11.8)**	-1.6 (2.8)
	Male	19	42.4 (10.2)	47.4 (12.9)	5.0 (13.8)	17	46.0 (13.1)	48.9 (11.1)	3.0 (11.7)	
STEM	No[§]	20	36.5 (10.8)	42.8 (10.9)	6.3 (9.6)**	17	41.4 (12.1)	52.2 (13.2)	10.8 (8.0)***	-4.1 (3.0)
	Yes	37	41.1 (11.3)	45.9 (12.9)	4.8 (14.4)*	32	43.0 (10.7)	45.1 (11.8)	2.1 (12.4)	
Fin. Aid	No[§]	9	38.0 (11.9)	41.0 (10.2)	2.9 (13.3)	21	41.8 (10.8)	52.3 (12.1)	10.4 (8.9)***	-2.2 (3.2)
	Yes	48	39.8 (11.2)	45.5 (12.5)	5.8 (12.9)**	28	42.9 (11.5)	44.1 (12.0)	1.2 (12.3)	
GPA	<3.0[§]	20	36.1 (12.1)	41.2(12.7)	5.1 (14.0)	17	44.2 (13.5)	43.9 (12.6)	-0.3 (14.0)	3.7 (2.8)
	>=3.0	31	40.7 (11.4)	46.2 (12.6)	5.6 (13.1)*	25	41.8 (10.1)	51.2 (13.2)	9.4 (9.3)***	

[‡]Regression coefficient (SD) from simple linear regression for score change (posttest – pretest) on course type is 0.18 (2.40). [§]Reference category. [§§]Obtained from multiple linear regression for score change on five covariates.

Impacts on Data Science Awareness, Aspiration, and Readiness among Students

Factor analysis was used to summarize students' responses to the 7 DS awareness and aspirations items in two constructs: "awareness" and "aspiration", as shown in Table 2. Consequently, we calculated the DS awareness and aspirations scores for each student by averaging the student's responses to the 3 awareness items and the 4 aspirations items, respectively. Similarly, we computed the DS readiness score from the 5 readiness items shown in Table 3. Although the factor analysis revealed that these 5 items load on two sub-constructs of readiness: general and R-related, the two sub-constructs were strongly correlated and hence were combined into one "readiness" construct.

Table 2. Results of the factor analysis on the DS awareness and aspiration items.

	Pre-survey		Post-survey	
Item	Aware.	Aspirat.	Aware.	Aspirat.
1. Are you aware that NCA&T offers courses in Data Science?§	0.70		0.87	
2. Are you aware that NCA&T offers an undergraduate certificate in Data Science?§	0.78		0.99	
3. Are you aware that NCA&T offers degrees with concentration in in Data Science?§	0.88		0.81	
4. Do you plan to take Data Science course(s) during your undergraduate program or during your graduate study (if you plan to do graduate studies)?‡		0.56		0.84
5. Do you plan to complete a certificate in Data Science during your undergraduate program or during your graduate study (if you plan to do graduate studies)?‡		0.80		0.86
6. Do you plan to complete a minor in Data Science during your undergraduate program or during your graduate study (if you plan to do graduate studies)?‡		0.87		0.83
7. Do you plan to complete a degree in Data Science during your undergraduate program or during your graduate study (if you plan to do graduate studies)?‡		0.80		0.86
Cronbach Alpha	0.82	0.83	0.92	0.91
Lewis-Tucker Index of Factoring Reliability	0.80		0.71	

§Possible responses were coded as "Yes=1" and "No=0". ‡Possible responses were coded as "Yes=2", "Not sure=1", and "No=0". The correlation between the two factors is 0.02 and 0.04 for pre-survey and post-survey.

Table 3. Results of the factor analysis on the DS readiness items.

	Pre-survey		Post-survey	
Item	Read. (General)	Read. (R)	Read. (General)	Read. (R)
1. I feel confident summarizing data sets using summary statistics and graphics in RStudio.§	0.95		0.93	
2. I feel confident performing basic statistical inference in RStudio§	0.89		0.94	
3. I feel confident performing basic modeling (linear and/or logistics regression).§		0.83		0.93
4. I feel confident creating reproducible data analysis reports in RStudio using R Markdown.§	0.84		0.84	
5. I feel adequately prepared to apply statistical and data-analytical techniques/tools to study a given topic.§		0.79		0.61
Cronbach Alpha	0.93	0.80	0.94	0.82
Lewis-Tucker Index of Factoring Reliability	1.00		0.94	

§Possible responses were reported on a 6-point Likert scale: "Strongly disagree=1" to "Strongly agree=6". The correlation between the two readiness sub-scales is 0.66 and 0.67 for pre-survey and post-survey, respectively.

Table 4 summarizes students' DS awareness, aspirations, and readiness scores in the DS-infused versus traditional sections of Intro Stats. Overall, students in the DS-infused sections had significantly greater gains in their DS awareness than students in the traditional sections (0.49 vs 0.30;

diff. in means = 0.19; p-value = 0.0393). This finding was further confirmed by the regression results that accounted for several demographic and academic factors (course type coeff. = 0.25; p-value = 0.0175). An equally important observation from Table 4 is that the DS-infused course seems to help close the DS awareness gaps among females and males; STEM and non-STEM students; financial aid and non-financial aid students; and students with varying academic preparations and achievement levels, as seen from comparing the pre- and post-survey means for these groupings. Considering students' self-reported readiness levels for DS, our results again suggested significantly higher positive impacts for the DS-infused course over the traditional course (1.30 vs 0.82; diff. in means = 0.48; p-value = 0.0442). Although the course effect became insignificant in the regression model accounting for the demographic and academic factors, it is quite encouraging to note the substantial increases in the readiness levels among female and non-STEM students in the DS-infused sections. Finally, financial aid students had significantly lower gains in readiness, especially in the traditional sections. Conversely, in both the DS-infused and traditional sections, we observe marginal drops in students' DS aspiration scores in the post-survey relative to the pre-survey. Although the drops are not statistically significant in both course types, the drops seem to be more prevalent in the DS-infused sections which suggests that infusing computing in the Intro Stats course seems to drive some students, especially females and lower achievers, away from aspiring for further DS education. This is consistent with findings in the literature that hinted at the difficulties associated with integrating computing into introductory courses (e.g., Woodard and Lee, 2021).

Table 4. Levels of DS awareness, aspiration, and readiness in two versions of the Intro Stats course.

		DS-Infused				Traditional§				
		n	Pre mean (SD)	Post mean (SD)	Mean diff (SD)	n	Pre mean (SD)	Post mean (SD)	Mean diff (SD)	Coeff. (SD)§§
		Awareness								
Course Type‡		45	0.18 (0.33)	0.67 (0.46)	0.49 (0.44)	49	0.10 (0.27)	0.40 (0.46)	0.30 (0.44)	0.25 (0.10)*
Sex	Female§	31	0.11 (0.26)	0.67 (0.46)	0.56 (0.45)	32	0.05 (0.21)	0.33 (0.45)	0.28 (0.43)	-0.08 (0.11)
	Male	14	0.33 (0.41)	0.67 (0.47)	0.33 (0.39)	17	0.18 (0.36)	0.51 (0.46)	0.33 (0.46)	
STEM	No§	13	0.05 (0.13)	0.62 (0.51)	0.56 (0.48)	20	0.12 (0.29)	0.42 (0.48)	0.30 (0.47)	0.02 (011)
	Yes	32	0.23 (0.37)	0.69 (0.44)	0.46 (0.43)	29	0.08 (0.26)	0.38 (0.44)	0.30 (0.42)	
Fin. Aid	No§	7	0.38 (0.45)	0.86 (0.38)	0.48 (0.47)	19	0.07 (0.21)	0.53 (0.46)	0.46 (0.49)	-0.18 (0.12)
	Yes	38	0.14 (0.30)	0.63 (0.46)	0.49 (0.44)	30	0.11 (0.31)	0.31 (0.44)	0.20 (0.38)	
AP Stat	No§	37	0.17 (0.33)	0.65 (0.46)	0.48 (0.44)	30	0.09 (0.28)	0.32 (0.44)	0.23 (0.40)	0.11 (0.12)
	Yes	8	0.21 (0.35)	0.75 (0.46)	0.54 (0.47)	19	0.11 (0.27)	0.51 (0.46)	0.40 (0.48)	
GPA	<3.0§	17	0.10 (0.26)	0.63 (0.48)	0.53 (0.47)	18	0.07 (0.24)	0.26 (0.42)	0.19 (0.38)	0.03 (0.10)
	>=3.0	23	0.23 (0.35)	0.74 (0.43)	0.51 (0.43)	25	0.13 (0.32)	0.53 (0.46)	0.40 (0.47)	
		Aspiration								
Course Type‡		47	0.37 (0.51)	0.29 (0.50)	-0.07 (0.55)	52	0.27 (0.39)	0.30 (0.50)	0.03 (0.51)	-0.14 (0.11)
Sex	Female§	31	0.30 (0.50)	0.15 (0.31)	-0.15 (0.48)	33	0.26 (0.39)	0.22 (0.42)	-0.04 (0.45)	0.15 (0.11)
	Male	16	0.52 (0.52)	0.58 (0.68)	0.06 (0.66)	19	0.30 (0.41)	0.45 (0.59)	0.15 (0.59)	
STEM	No§	13	0.17 (0.33)	0.09 (0.24)	-0.08 (0.43)	21	0.27 (0.43)	0.23 (0.42)	-0.05 (0.53)	0.01 (0.12)
	Yes	34	0.45 (0.55)	0.38 (0.56)	-0.07 (0.60)	31	0.27 (0.38)	0.36 (0.54)	0.08 (0.49)	
Fin. Aid	No§	7	0.61 (0.61)	0.54 (0.44)	-0.07 (0.40)	20	0.39 (0.46)	0.34 (0.45)	-0.05 (0.57)	0.11 (0.14)
	Yes	40	0.33 (0.49)	0.26 (0.51)	-0.08 (0.58)	32	0.20 (0.34)	0.28 (0.53)	0.08 (0.46)	
AP Stat	No§	38	0.38 (0.51)	0.27 (0.45)	-0.11 (0.43)	32	0.29 (0.41)	0.27 (0.44)	-0.03 (0.36)	0.18 (0.13)
	Yes	9	0.33 (0.56)	0.42 (0.71)	0.08 (0.94)	20	0.24 (0.37)	0.36 (0.58)	0.13 (0.68)	
GPA	<3.0§	17	0.37 (0.56)	0.16 (0.39)	-0.21 (0.49)	18	0.15 (0.30)	0.28 (0.58)	0.13 (0.58)	-0.07 (0.12)
	>=3.0	25	0.45 (0.51)	0.33 (0.47)	-0.12 (0.42)	27	0.34 (0.43)	0.27 (0.44)	-0.07 (0.49)	

		Readiness								
Course Type‡		35	2.32 (0.80)	3.62 (1.08)	1.30 (1.06)	32	2.04 (0.72)	2.86 (1.07)	0.82 (0.84)	0.37 (0.24)
Sex	Female§	23	2.10 (0.72)	3.57 (1.06)	1.47 (0.95)	18	2.03 (0.60)	2.67 (1.04)	0.63 (0.90)	-0.07
	Male	12	2.75 (0.79)	3.73 (1.16)	0.98 (1.24)	14	2.06 (0.88)	3.11 (1.09)	1.06 (0.73)	(0.24)
STEM	No§	9	2.20 (0.71)	3.82 (0.95)	1.62 (0.66)	10	2.00 (0.55)	2.92 (0.94)	0.92 (0.91)	-0.07
	Yes	26	2.36 (0.84)	3.55 (1.13)	1.19 (1.16)	22	2.06 (0.80)	2.84 (1.15)	0.77 (0.83)	(0.27)
Fin. Aid	No§	6	2.10 (0.72)	3.57 (1.34)	1.47 (1.12)	12	2.07 (0.40)	3.42 (1.10)	1.35 (1.05)	-0.75
	Yes	29	2.37 (0.82)	3.63 (1.05)	1.27 (1.07)	20	2.03 (0.87)	2.53 (0.93)	0.50 (0.48)	(0.28)**
AP Stat	No§	28	2.31 (0.87)	3.64 (1.05)	1.32 (1.01)	20	2.01 (0.69)	2.93 (1.28)	0.92 (0.97)	-0.45
	Yes	7	2.34 (0.43)	3.57 (1.30)	1.23 (1.33)	12	2.10 (0.81)	2.75 (0.63)	0.65 (0.56)	(0.28)
GPA	<3.0§	11	2.24 (0.66)	3.53 (0.85)	1.29 (0.89)	11	1.85 (0.90)	2.67 (1.34)	0.82 (0.86)	-0.13
	>=3.0	19	2.35 (0.95)	3.44 (1.15)	1.09 (0.99)	18	2.19 (0.60)	3.02 (0.93)	0.83 (0.91)	(0.25)

‡Regression coefficient (SD) from simple linear regression for score change (post– pre) on course type (DS-Infused vs Traditional) is 0.19 (0.09)*, -0.10 (0.11), and 0.48 (0.24)* for awareness, aspirations, and readiness, respectively. §Reference category. §§Obtained from multiple linear regression for score change on six covariates.

DISCUSSION

Motivated by the widespread calls for integrating computing in the Statistics curriculum to better prepare our graduates for the workplace and the growing need for graduates with data analytical skills, this paper proposed and evaluated a DS-infused Intro Stats course design. The preliminary results drawn from data collected at a large HBCU institution in the Fall of 2022 indicated that the DS-infused design can significantly enhance students' levels of DS awareness and readiness. The efficacy of this design in boosting students' statistical gains is yet to be confirmed from further data collection. On the contrary, infusing computing in Intro Stats seemed to drive some students away from aspiring for further DS education. This is in line with other findings in the literature that hinted at the complexity of computing and the challenges of integrating computing into introductory courses.

ACKNOWLEDGMENT

The authors thank Prof. Nathan Tintle for sharing their revised version of the CAOS test. This work was supported by the NSF Directorate for STEM Education under grant No. 2106945.

REFERENCES

Cobb, G. (2015). Mere renovation is too little too late: we need to rethink our undergraduate curriculum from the ground up. *The American Statistician*, 69, 266-282.

delMas, R.C., 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.

GAISE College Report ASA Revision Committee (2016). *Guidelines for Assessment and Instruction in Statistics Education College Report*. http://www.amstat.org/education/gaise

Horton, N.J., Baumer, B.S. & Wickham, H. (2015). Setting the stage for data science: integration of data management skills in introductory and second courses in statistics. *CHANCE*, *28*(2):40-50.

Horton, N.J. and Hardin, J.S. (2021). Integrating computing in the statistics and data science curriculum: Creative structures, novel skills and habits, and ways to teach computational thinking. *Journal of Statistics and Data Science Education*, 29:sup1 S1-S3.

Nolan, D., & Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician*, 64, 97–107.

Tintle, N., Clar, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T. & Vanderstoep, J. (2018). Assessing the association between precourse 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.

Woodard, V. & Lee, H. (2021). How students use statistical computing in problem solving. *Journal of Statistics and Data Science Education* 29(1), 1– 18.