USING A MIDTERM WARNING SYSTEM TO IMPROVE STUDENT PERFORMANCE AND ENGAGEMENT IN AN INTRODUCTORY STATISTICS COURSE: A RANDOMIZED CONTROLLED TRIAL

NOOSHIN KHOBZI ROTONDI
Ontario Tech University
nooshin.rotondi@ontariotechu.ca

DAVID RUDOLER
Ontario Tech University
Ontario Shores Centre for Mental Health Sciences
david.rudoler@ontariotechu.ca

WILLIAM HUNTER
Ontario Tech University
bill.hunter@ontariotechu.ca

OLAYINKA SANUSI
Ontario Tech University
olayinka.sanusi1@ontariotechu.net

CHRIS COLLIER
Ontario Tech University
chris.collier@ontariotechu.net

MICHAEL ANTHONY ROTONDI
York University
mrotondi@yorku.ca

ABSTRACT

This article reports on an evaluation the effectiveness of e-mailed grade "nudges" on students' performance and engagement in an introductory statistics course for undergraduate health science students. In 2020–2021, 358 students were randomized to an e-mail (n = 178) or no e-mail (n = 180) group. The intervention e-mail contained information on each student's predicted final grade (grade nudge). Using two-sample t-tests, the statistical analysis of final grades in the course showed a higher compatibility with a model of no mean difference for students in the e-mail (73.5%) vs. no e-mail (72.1%) group. Comparison of the distributions of final grades between the two groups, however, suggested the e-mailed nudges may be related to slight improvements in final grades. Specifically, the median final grade was higher in the e-mail group (74.6 vs. 72.4); the Q1 value in the e-mail group was also higher, and the interquartile range was similar: no e-mail group (15.8) vs. e-mail group (14.2). Students also completed the Scale of Student Engagement in Statistics (SSE-S). Total engagement, affective and cognitive subscale scores of the SSE-S were higher in the e-mail group, resulting in low compatibility with a model of no difference in engagement scores. Overall, the results showed there is potential for our midterm warning system to be used to improve outcomes, particularly given that it is simple to implement, cost-effective, and easily scalable.

Keywords: *Statistics education research; Student engagement; e-mail nudges; Introductory statistics; Randomized controlled trial*

Statistics Education Research Journal, 22(3), Article 2. https://doi.org/10.52041/serj.v22i3.406
© International Association for Statistical Education (IASE/ISI), November 2023

1. INTRODUCTION

Mandatory undergraduate statistics courses, including at the introductory level, can be challenging for many students in non-statistics majors. Typically, students who take these courses have little or no prior knowledge of statistics, and most will not pursue statistics beyond the limited course requirements of their majors (e.g., Life Sciences or Health Sciences). Engaging students within these contexts and supporting academic success can be challenging when students do not perceive the course as being directly related to their future career aspirations. This is a particular issue given evidence that student engagement is a key contributor to student success (Kahu, 2013). Within STEM (science, technology, engineering, and mathematics) courses, interest in STEM itself is an important indicator of success, as are related factors such as the desire to engage with STEM content, student beliefs that the course material is relevant, and interest in STEM careers (Crisp et al., 2009; Jones et al., 2010; Tai et al., 2006). These factors may be lacking in students who are required to take introductory statistics as part of their majors. Instructors of introductory statistics courses may be able to improve student outcomes by creating a learning environment that is both motivating and engaging.

There is limited research on effective educational interventions to improve student engagement and academic success in introductory statistics courses at the undergraduate level. In science and economics courses, interventions typically include providing information about students' standing in the course and "nudging" them to use additional support services (Moss & Yeaton, 2015). Researchers have found that reminders to students about their standing in the course can improve student performance (Chen & Okediji, 2014; Smith et al., 2018). Smith et al. (2018) performed a randomized trial in which students were reminded of their current grade on their homework assignments via e-mail. The intervention led to an improvement in homework performance by four percentage points. In a systematic review on nudges in education, Damgaard and Nielsen (2018) concluded that "few interventions produce positive effects for everyone, and some nudges even have negative effects" (p. 314). In the context of nudging students to access resources there is research, however, suggesting that these increase students' use of services (Butcher & Visher, 2013; Pugatch & Wilson, 2018). Specifically, Pugatch and Wilson (2018) who used a randomized experiment informing students about peer tutoring services via a postcard, found an increase in attendance but no change in performance.

E-mailed "grade nudges" have also been used to improve the study habits of lower-performing students. One study used an intervention consisting of personalized e-mails wherein the instructor expressed concerns about the student's course standing, and provided specific study advice (Deslauriers et al., 2012). Relative to those who did not receive the intervention, students who received personalized e-mails obtained higher mean scores on their second midterm compared to their first term test. Similarly, Gordanier et al. (2019) provided students with poor performance and excessive absences in large undergraduate economics courses, information on their standing. Students were also referred to the university's Student Success Center for additional academic support. The researchers found that the intervention improved student scores on the final exam by 6.5–7.5 percentage points (Gordanier et al., 2019). This result was corroborated in a descriptive survey on undergraduate students at a large university. The survey results revealed that students who engaged more with the success center earned significantly higher grades than those who visited the center occasionally (Osborne et al., 2019). Similarly, Boretz (2014) found that academic and learning support programs were associated with a lower proportion of students who received an unsatisfactory grade, and improved engagement in first-year college students.

Academic achievement is closely linked with student engagement (Appleton et al., 2008; Carini et al., 2006; Willingham et al., 2002). The construct of engagement, though not consistently operationalized, has been shown to be multidimensional, consisting of three main factors (Fredricks & McColskey, 2012; Jimerson et al., 2003; Marks, 2000; Whitney et al., 2019): behavioral (e.g., participation/interaction), affective (e.g., interest in learning) and cognitive (e.g., higher order thinking beyond a "general curiosity in statistics" [Whitney et al., 2019, p. 558]). As stated by Whitney et al. (2019), "… student engagement is relatively malleable … and is therefore an important target for intervention" (p. 553). To our knowledge, there are no studies measuring engagement in undergraduate students using instruments with appropriate evidence of validity designed specifically for introductory statistics courses. Lawton and Taylor (2020), however, used a two-item daily engagement survey to collect data on students' perception of their engagement in an introductory course. The surveys revealed

a moderate level of engagement, with variability attributed to the course content and in-class activities. Another study conducted by Muir et al. (2020), measured academic engagement using an adapted version of the 9-item Utrecht Work Engagement Scale (Schaufeli et al., 2006). Using an experimental crossover design, the authors reported that online student response systems may be appropriate tools to increase student engagement in undergraduate statistics courses.

Overall, results from previous educational studies highlight the importance of implementing early intervention programs to improve student performance and engagement. There is limited empirical evidence, however, that shows the efficacy of related educational interventions at an undergraduate level in introductory statistics. To address this gap, we aim to evaluate the effects of a midterm warning system (e-mailed grade nudges) on students' overall performance and engagement in a mandatory introductory statistics course for health science students within a midsized Canadian university.

2. METHODS

2.1. COURSE DESCRIPTION AND SAMPLE

The undergraduate statistics course, Introduction to Statistics for Health Sciences, is required for all undergraduate students registered in the Faculty of Health Sciences. The programs and discipline are diverse, including nursing, kinesiology, public health, human health science, medical laboratory science, and allied health science. The introductory course emphasizes critical appraisal skills in assessing evidence presented in health sciences, with a focus on real-life relevance. The application of statistical methods to the study of research questions is explored in terms of both descriptive and inferential statistics. Students are taught to perform calculations by hand and using *SPSS*, as well as how to interpret the results of statistical analyses. The course is typically taught in-person, but it was offered online (asynchronous) with synchronous virtual tutorials for the 2020–2021 academic terms due to the COVID-19 pandemic. At the time of the study, the lead author was the instructor for all sections of the course, with a typical enrolment of 200–250 students per term. Detailed course information is provided in Table 1.

Table 1. Detailed course information including topics covered and timing of instrument completion, course assessments, and e-mailed nudges

Week	Topic	Notes
1	Introduction: Course Information, Populations/Samples, Descriptive/inferential statistics, Measurement scales, Study designs	
2	Intro to SPSS, Describing and Exploring Data, Measures of Central Tendency	
3	Measures of Variability, Simple Probability and Binomial Distribution	Completion of baseline questionnaire
4	Normal Distribution and z-scores Standard Error of the Mean, Confidence Intervals	
5	Article Review Assignment 1 introduced Study Week (No classes)	Test #1 Study Week (No classes)
6	Logic of Hypothesis Testing One Sample Inference (Z-test and t-test)	Assignment #1 Due
7	Two Sample Inference (unpaired and paired data)	E-mail group receives intervention e-mail
8	Analysis of Categorical Data (chi-square tests), Review for Test #2	
9	Statistical significance vs. clinical importance, How to write Methods and Results sections	Test #2 Completion of Scale of Student Engagement in Statistics
10	ANOVA	
11	Correlation and regression	Assignment #2 Due Completion of Scale of Student Engagement in Statistics
12	Review for Final Exam	

In the 2020–2021 academic year, four sections of the course were offered over two terms. In the Fall and Winter, 158 and 200 students were registered in the introductory statistics course, respectively. All 358 students were included in the study. Given that our focus was entirely on improvement of student performance and engagement in the course, the university's Research Ethics Board (REB) approved our request for a waiver of consent and provided an exemption from REB review (File #15936).

2.2. STUDY MEASURES

We developed a brief demographic questionnaire that was administered to students online using *Google Forms*. The questionnaire items can be found in Table 2 (see Section 2.3 Study Design). Our primary outcome, student performance, was measured using final grades (percentages). Our secondary outcome of interest was student engagement. This was measured using the Scale of Student Engagement in Statistics (SSE-S), which consists of 24 Likert-type items, with eight items each reflecting the affective, behavioral, and cognitive factors of engagement (Ober et al., 2021; Whitney et al., 2019). Responses were provided using a 5-point Likert type scale indicating the extent to which participants agreed with the statement (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree). The SSE-S has demonstrated good internal consistency with Cronbach's alpha scores of 0.88, 0.89, 0.85, and 0.79 for the full-scale score, affective, behavioral, and cognitive subscale scores, respectively (Whitney et al., 2019). Similarly, the test–retest reliability of the full and subscale scores are very high: 0.86 for the full-scale score, and 0.85, 0.85, and 0.77, respectively, for the affective, behavioral, and cognitive subscale scores (Whitney et al., 2019).

2.3. STUDY DESIGN

All students were invited to complete the baseline questionnaire in Week 3. To minimize potential conflict given the lead author's dual role as instructor and study lead, randomization was performed by the study co-lead. The student population of the Faculty of Health Sciences is diverse with a high level of variability in terms of interest in statistics, existing mathematics or statistics knowledge, and overall student performance in the university, amongst other factors. A randomized controlled trial (RCT) study design was selected to ensure the intervention and control groups were balanced on potential confounders, thereby contributing to unbiased estimates of the intervention effect. There is, however, the potential for attrition bias (differential loss of participants). To avoid a substantial decrease in power, we enrolled all students from the four course sections into the study to compensate for expected withdrawals. Additional details about the potential impact of student withdrawals can be found in the Discussion (Section 4).

Students were randomized to an intervention/e-mail group vs. no e-mail group (control); the former received an e-mail at the end of Week 7, following the completion of the first test and assignment in the course. The e-mail messages included information on the students' predicted final grade in the course: the basis for prediction is explained in Section 2.4 Statistical Methods. For lower-performing students (predicted final grade of C+ or lower), we also provided information on academic resources available at the university through a link to a Google Doc. This document included information on different types of services and how students may access them, including one-on-one mathematics support, statistics workshops, and study skills workshops. Information about the same resources was made available to all students in Week 2, as part of standard practice in the course. The e-mail messages for the intervention group were tailored to each predicted grade level (A+ to F). Briefly, students predicted to receive: 1) A- or higher were encouraged to keep up their efforts; 2) B- to B+ were reminded about their potential for improvement but also congratulated for their efforts; and 3) C+ or lower were gently reminded of the consequences of low grades and encouraged to use academic resources.

Students in both the e-mail vs. no e-mail groups were invited to complete the SSE-S online through *Google Forms* at the start of Weeks 9 and 11 to allow for assessment of test-retest reliability in the undergraduate university population. Students were given five days to complete the scale from the time the invitations were e-mailed to them. The timing of instrument completion, course assessments and e-mailed nudges are listed in Table 1.

Baseline characteristics. Of 358 students initially included in this study, 324 completed the demographic questionnaire. The comparison of baseline characteristics between the e-mail and no e-mail groups is provided in Table 2. The comparison indicated no apparent differences between the groups.

Table 2. Baseline characteristics at randomization for all students who completed the demographic questionnaire (n = 324)

Student Characteristics	E-mail Group n = 160	No E-mail Group n = 164	p-value[d]
Age in years, mean (SD)	22.1 (4.9)	21.2 (5.2)	0.1309
Gender, *n* (%)			
Female	123 (76.9)	113 (68.9)	0.1067
Male	37 (23.1)	51 (31.1)	
Program of study[a], *n* (%)			
Kinesiology	55 (34.6)	51 (31.3)	0.9661
Nursing	33 (20.7)	36 (22.1)	
Human Health Science	24 (15.1)	25 (15.3)	
Public Health	24 (15.1)	24 (14.7)	
MLSc or AHSc	23 (14.5)	27 (16.6)	
Current status, *n* (%)			
Full time	151 (94.4)	156 (95.7)	0.5816
Part time	9 (5.6)	7 (4.3)	
Cumulative GPA, *n* (%)			
3.7+ (A)	36 (22.5)	38 (23.2)	0.9780
2.7 to < 3.7 (B)	89 (55.6)	88 (53.7)	
< 2.7 (C – F)	24 (15.0)	25 (15.2)	
Don't know/Prefer not to answer	11 (6.9)	13 (7.9)	
Year of study, *n* (%)			
Second Year	106 (66.2)	112 (68.7)	0.6368
Other[b]	54 (33.8)	51 (31.3)	
Last math course, *n* (%)			
University	83 (52.2)	85 (52.2)	0.3147
College	18 (11.3)	11 (6.7)	
High school/Elementary school	58 (36.5)	67 (41.1)	
Hours worked for pay in typical week, *n* (%)			
0	53 (34.6)	54 (33.3)	0.6692
1-10	21 (13.7)	26 (16.0)	
11-20	38 (24.8)	45 (27.8)	
21-30	23 (15.0)	16 (9.9)	
31+	18 (11.8)	21 (13.0)	
Recent immigrant[c], *n* (%)			
Yes	7 (4.4)	< 5	0.3360
No	153 (95.6)	160 (97.6)	
Language first learned in childhood and still understand			
English	107 (67.7)	115 (71.0)	0.5263
Other	51 (32.3)	47 (29.0)	

Note. Total sample of 358 students; at randomization, 178 were assigned to the e-mail group and 180 to the no e-mail group. Differences between e-mail and no e-mail groups were tested using Chi-square tests for all variables, except for Age (t-test). [a]For Program of Study, MLSc = Medical Laboratory Science and AHSc = Allied Health Science. [b]The 'Other' category in Year of Study included first-, third-, fourth-, and fifth-year students. [c]Recent immigrants are defined as those who settled in Canada less than 5 years ago. [d]Raw *p*-values obtained from independent tests.

2.4. STATISTICAL METHODS

Predicting final grades. By mid-semester, students in the course completed a test and an assignment. Predicted grades for students in the e-mail group were based on the grades of 565 former

students who were all taught by the lead author in the 2019–2020 academic terms. The previous students' marks for Test 1, Assignment 1, and their final course grades were used to obtain regression coefficients. These coefficients were applied in a linear prediction model containing Test 1 and Assignment 1 marks for students in the e-mail group, allowing for the prediction of final grades.

The regression model for final grade prediction was:

$$\textit{Predicted final grade} = 23.3835 + 0.4610\textit{Test1} + 0.2327\textit{Assign1}$$

where *Test1* denotes the students' percentage grade on Test 1 and *Assign1* denotes the students' percentage grade on Assignment 1. Note that only a handful of the 565 former students achieved grades in the high 90s, and for this reason the highest predicted grade using this model is approximately 93%. Furthermore, based on the regression coefficients in our equation, it is apparent that Test 1 has the largest impact on predicting final grades. This corresponds to the assessments' relative weights in the course (15% for Test 1, and 10% for Test 2) and the potential importance of test performance on final exam scores, which are highly weighted in this course (30%).

SSE-S: Scoring, Internal Consistency and Test-Retest Reliability. Negatively keyed items in the SSE-S were reverse-coded. Among completed scales, missing items were estimated using mean imputation. A full-scale score was computed by averaging scores of all 24 items. The three subscale scores were also created by averaging the item scores within each subscale. These analyses were performed using *SAS*® Version 9.4.

Across both study groups, 60 students were missing at least one entire scale; of these, 26 were missing the scale at both Weeks 9 and 11, 17 students were missing the SSE-S in Week 9, and 17 were missing the scale in Week 11. Of the 332 students who had completed at least one entire scale, 20 were missing one item, four were missing two items, and three were missing three items. The specific items missing varied, which suggested random missingness.

Among students who completed the demographic questionnaire (n = 324), 34 students submitted questionnaires with at least one scale missing. In terms of group assignment (e-mail vs. no e-mail group) and most baseline characteristics (results not shown), the comparison of students missing at least one scale to those who completed both scales indicated that results were more compatible with a model of no difference. The results, however, were less compatible with a model of no difference when testing the association between reported cumulative GPA (cGPA) and missing at least one scale. Specifically, a higher proportion of students with cGPAs < 2.7 (C to F) were missing at least one scale, whereas a lower proportion of students with cGPAs of 3.7+ (A- or higher) were missing at least one scale.

Internal consistency was assessed using Cronbach's alpha and corresponding 95% compatibility intervals (*CI*s; Rafi & Greenland, 2020) for the full- and subscale scores based on the intraclass correlation coefficient (*ICC*) two-way mixed model (Baumgartner & Chung, 2001; Bravo & Potvin, 1991) in SPSS (Version 27). For test-retest reliability, we calculated the *ICC*s and 95% *CI*s for the full- and subscale scores using *SPSS* (Version 27), based on a two-way mixed effects model with an absolute agreement definition for single raters (Koo & Li, 2016; Shrout & Fleiss, 1979).

The SSE-S had high levels of internal consistency over the entire 24-item scale, Cronbach's α = 0.90, 95% *CI* [0.89, 0.92], as well as within each 8-item subscale, affective α = 0.92, 95% *CI* [0.91, 0.94], behavioral α = 0.83, 95% *CI* [0.81, 0.86], and cognitive α = 0.79, 95% *CI* [0.75, 0.82]. Among the 298 students who completed the SSE-S at both Weeks 9 and 11, the test-retest reliability of the full- and sub-scale scores were also high. For the full-scale, *ICC* = 0.85, 95% *CI* [0.81, 0.87]; affective sub-scale, *ICC* = 0.84, 95% CI [0.80, 0.87]; behavioral subscale, *ICC* = 0.80, 95% CI [0.75, 0.84]; and cognitive sub-scale, *ICC* = 0.71, 95% *CI* [0.65, 0.77], respectively.

Testing the effectiveness of the midterm warning system (E-mail intervention). We assessed the differences in 1) final grades (percentages), and 2) engagement scores (full- and subscales) between the e-mail and no e-mail groups using two-sample t-tests. For the comparison of full- and subscale scores between the e-mail vs. no e-mail groups, we calculated Cohen's *d* effect sizes and 95% *CI*s using the corresponding means, standard deviations, and group sizes in the *R* package "metafor" (Viechtbauer, 2010) using the statistical programming language, *R 4.0.1*. The analysis was completed only for the

SSE-S scores obtained at Week 11, three weeks after the e-mails were sent to the intervention group. This allowed for a reasonable test of effectiveness, given that lower performing students in the e-mail group would have more than one week to seek academic support and resources.

3. RESULTS

3.1. EFFECTIVENESS OF THE MIDTERM WARNING SYSTEM (E-MAIL NUDGE INTERVENTION)

Final grades. Of the 358 students in this study, 22 were missing final grade data (14 dropped the course and 8 requested exam deferrals). In the analysis of final grades of the remaining 336 students (Table 3), the comparison of e-mail group vs. no e-mail group indicated the results were more compatible with a model of no difference.

Table 3. Comparison of final grades between e-mail and no e-mail groups (n = 336)

Outcome	E-mail Group Mean (*SD*) *n* = 161	No E-mail Group Mean (*SD*) *n* = 175	Mean Difference, [95%, *CI*]
Final grade (%)	73.5 (11.1)	72.1 (10.9)	1.3, [-1.0, 3.7]

Note. Missing final grade data: *n = 22*

In addition, we examined the distribution of final grades across the two groups (Figure 1). The median final grade was higher in the e-mail group (74.6 vs. 72.4), and the interquartile range was similar: no e-mail group (15.8) vs. e-mail group (14.2). Overall, the final grades were generally distributed in a similar manner. The lower median in the no e-mail group, and the higher Q1 value in the e-mail group, suggested the intervention e-mail may be related to slight improvements in final grades, despite the greater compatibility with a model of no difference (Table 3).

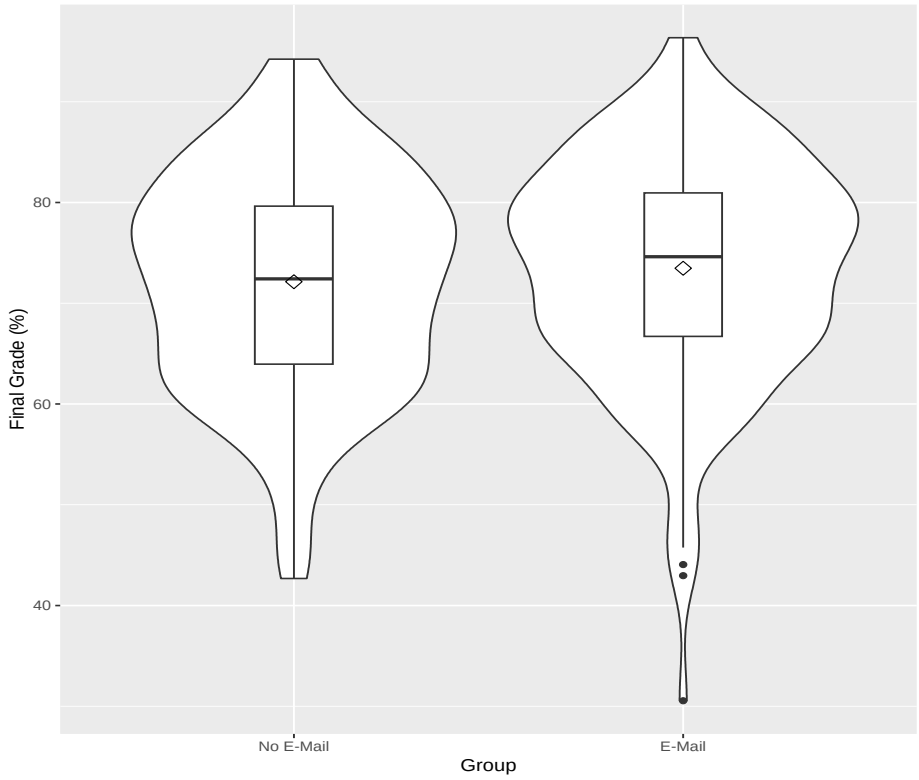

Figure 1. Comparison of distributions of final grades between e-mail and no e-mail groups (n = 336).

Furthermore, we performed an exploratory subgroup analysis (Figure 2) to investigate potential differences in the behavior of lower performing (C+ or lower) students (n = 132) compared to higher performing (B- or higher) students (n = 204). For the latter, the results were similar to those of the overall group, though we observed a smaller difference in median final grades between the e-mail and no e-mail groups. Within the lower performing students, the median final grade was over two percent higher in the e-mail group. The distribution of grades was more negatively skewed in the e-mail group, likely due to one extremely low outlier value (Figure 2). Also, a larger proportion of students in the e-mail group earned higher final grades than the no e-mail group, as demonstrated by the wider section of the violin plot above the median. Overall, there was more variability in the final grades when comparing lower performing to higher performing students. As expected, all 22 students with missing final grade data were lower performing; however, 17 of these were in the e-mail group.

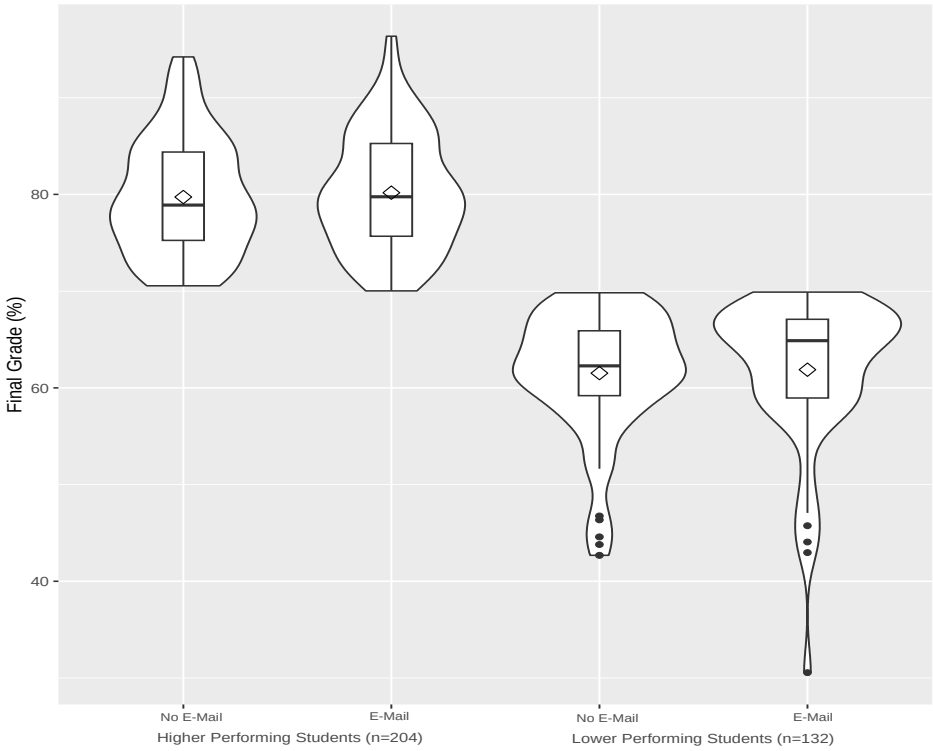

Figure 2. Comparison of distributions of final grades between lower and higher performing students.

Student engagement. Our results indicated that students in the e-mail group had higher engagement scores than those who did not receive the intervention e-mail. Specifically, our comparisons of the full-scale, and affective and cognitive subscales of the SSE-S completed at Week 11 (see Table 4) suggested low compatibility with a model of no difference in engagement scores. Using benchmarks suggested by Cohen (1988), the effect sizes would be considered between small and medium.

Table 4. Comparison of SSE-S full- and subscale scores between e-mail and no e-mail groups at Week 11 (n = 315)

SSE-S	E-mail Group Mean (*SD*) n = 153	No E-mail Group Mean (*SD*) n = 162	Effect size, [95%, *CI*]
Total engagement	85.8 (11.5)	82.1 (11.9)	0.32, [0.09, 0.54]
Affective	25.3 (5.8)	23.4 (6.4)	0.31, [0.09, 0.53]
Behavioural	31.4 (4.4)	30.9 (4.6)	0.11, [-0.11, 0.33]
Cognitive	29.1 (4.3)	27.8 (4.4)	0.30, [0.08, 0.52]

Note. Effect size is measured using Cohen's *d.*

4. DISCUSSION

Our randomized trial, based on health science students enrolled in a compulsory introductory, undergraduate statistics course suggested there is higher compatibility with a model of no difference in final grades for students in the e-mail vs. no e-mail group. The e-mailed nudges provided students in the intervention group with information about their predicted final grades, as well as reminders about the academic support services available at the university for lower performing students. Our finding is in contrast to previous studies that found providing information to students about their course standing and nudging them to access resources improved student performance (Chen & Okediji, 2014; Deslauriers et al., 2012; Gordanier et al., 2019; Smith et al., 2018). Comparing the distributions in final grades between the two groups, however, suggests the intervention e-mail may be related to slight improvements in final grades. Lower performing students may have driven this result, with the majority of those in the e-mail group earning higher final grades than the no e-mail group.

There may be several reasons for the discrepancy in our results compared to previous research: few studies used a RCT study design, all were based in different populations (e.g., undergraduate economics and science students), e-mail "nudges" varied in terms of the types of information provided, and student performance or success was measured differently across studies. To our knowledge, our study is the only one to provide students with a predicted final grade, rather than current standing in the course. While predicted final grades were based on students' current standing, this slight difference in the type of information provided may have contributed to our contrasting result. Indeed, it is possible that some students interpreted the prediction as in some way immutable. Furthermore, we evaluated improvements in final grades, whereas the outcomes of interest in previous studies included homework performance (Smith et al., 2018), Midterm 2 vs. Midterm 1 grades (Deslauriers et al., 2012), and final exam scores (Gordanier et al., 2019).

Another notable difference in our study design compared to previous studies is that we provided information on academic resources and how students may access them. The onus was therefore on each student to access the resources. In fact, some students concerned about their low predicted grades (n = 8) contacted the two leading authors to discuss their performance in the course and how they may improve, even though they were not required to do so. In other studies students were referred to academic support services (Gordanier et al., 2019), and provided specific study advice or asked to meet face-to-face with the instructor (Deslauriers et al., 2012). This may suggest a need for a more hands-on intervention, whereby lower-performing students are required to meet with the professor, or students are referred directly to the university academic support services. This was similar to survey data from a business school in the United States, which showed that students believed academic interventions were instrumental to academic success due to increased engagement and communication with advisors and instructors (Niranjan et al., 2015).

Despite the apparent ineffectiveness of our intervention in improving final grades, there is evidence indicating that students in the e-mail group accessed the Google Doc containing information on academic supports available at the university. By retrieving the view history in Google Docs, we found that 36.8% of students with a predicted grade of C+ or lower accessed the document. Unfortunately, we cannot determine how many of these students attended support services programmes or made use of the resources offered in the document. Other researchers, however, have found that informing students about peer tutoring services increased uptake of services, but did not always contribute to changes in performance (Pugatch & Wilson, 2018).

In our study, providing students with information about their predicted final grades and academic services available at the university was effective in increasing student engagement in the course. As shown in Table 4, total engagement, and affective and cognitive subscale scores were higher in the e-mail group, with effect sizes of 0.32, 0.31 and 0.30, respectively. While these values are considered small to medium (Cohen, 1988), others have noted that the benchmarks are arbitrary and should not be interpreted rigidly (Thompson, 2007). Indeed, even small effect sizes can have large consequences, as in our study where a simple, low-cost e-mail intervention contributed to improved engagement scores in all areas except for the behavioral subscale. Nonetheless, given the novelty of the SSE-S and lack of comparisons in the literature, applying these benchmarks to our study would be considered useful (Cohen, 1988).

Of final note is the apparent lack of improvement in behavioral engagement. It is possible that our intervention, while effective in improving affective factors (e.g., interest and motivation in learning statistics), and cognitive factors (e.g., ability to make connections between topics in the course and thinking in different ways to solve problems), may not have been sufficient to bring about actual change in behavior (e.g., studying for statistics on a regular basis and taking good notes on the material). Future studies may need to examine the impact of interventions that directly encourage change in behavior, such as one-on-one meetings with the professor and required attendance at academic support services or study skills workshops.

4.1. STRENGTHS AND LIMITATIONS

Our study has many strengths: 1) we used a robust RCT study design with a relatively large sample size; 2) the professor for all of the introduction to statistics classes was the same, thus we were able to control for content delivery directly through the study design; and 3) we used a scale with previous evidence of validity (SSE-S) designed to measure student engagement specifically in statistics courses. Nonetheless, there were some limitations to our study. All 22 students missing final grade data (due to dropping the course or requiring exam deferrals) were among the lower performing students (C+ or lower). This finding is not surprising as lower performing students are more likely to withdraw from the course, but over three times of the missing data were in the e-mail group (17 vs. 5). This may have contributed to an underestimation of the overall difference in mean final grades between the e-mail and no e-mail groups. Furthermore, the number of students dropping the course was much higher than previous terms, possibly due to the ongoing stresses of the COVID-19 pandemic during which this study took place. Students who were already lower performing and struggling due to the broad effects of COVID-19 may have been more adversely impacted by the e-mail nudges. The potential negative effects of nudges have been previously reported by others (Damgaard & Nielsen, 2018).

In terms of the SSE-S, the tool was developed for high school Advanced Placement (AP) students in statistics, which may not be appropriate for use in second- or third-year undergraduate students. While we found acceptable levels of internal consistency and high test-retest reliability values, the SSE-S may require further evidence of validity for use with university populations. Furthermore, our finding that students with lower cGPAs were less likely to complete the SSE-S points to possible non-ignorable missingness. This may have contributed to the higher engagement scores in the e-mail group, that is, the intervention improved engagement partly due to lower performing (and thus lower engaged students) not completing the SSE-S. Despite these limitations, the outcomes from the randomization of students to e-mail vs. no e-mail group provides robust evidence for the positive impact of our intervention on engagement in undergraduate students.

4.2. CONCLUSION

The midterm warning system of e-mailed grade nudges was effective in improving student engagement in a mandatory introductory statistics course. While we found no statistical evidence of effectiveness in terms of final grades, a comparison of the distributions between the two groups suggests the e-mailed nudges may have contributed to slight improvements in final grades for many students. These results indicate that there is potential in our midterm warning system, particularly given that it is simple to implement, cost-effective and easily scalable across similar courses in various post-secondary institutions. Nonetheless, we must also be mindful of potential adverse impacts given the observed higher percentage of course withdrawal in lower performing students who received the e-mailed nudges. Future research should include “enhanced” interventions that may directly influence students’ behavior, such as required meetings with the professor and attendance at academic support services.

REFERENCES

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools, 45*, 369–386. https://doi.org/10.1002/pits.20303

Baumgartner, T. A., & Chung, H. (2001). Confidence limits for intraclass reliability coefficients. *Measurement in Physical Education and Exercise Science*, *5*(3), 179–188. https://doi.org/10.1207/S15327841MPEE0503_4

Boretz, E. (2014). Midsemester academic interventions in a student-centered research university. *Journal of College Reading and Learning*, *42*(2), 90–108. https://doi.org/10.1080/10790195.2012.10850356

Butcher, K. F., & Visher, M. G. (2013). The impact of a classroom-based guidance program on student performance in community college math classes. *Educational Evaluation and Policy Analysis*, *35*(3), 298–323. https://doi.org/10.3102/0162373713485813

Bravo, G., & Potvin, L. (1991). Estimating the reliability of continuous measures with Cronbach's Alpha or the intraclass correlation coefficient: Toward the integration of two traditions. *Journal of Clinical Epidemiology*, *44*(4–5), 381–390. https://doi.org/10.1016/0895-4356(91)90076-L

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education*, *47*(1), 1–32. https://doi.org/10.1007/s11162-005-8150-9

Chen, Q., & Okediji, T.O. (2014). Incentive matters! The benefit of reminding students about their academic standing in introductory economics courses. *The Journal of Economic Education*, *45*(1), 11–24. https://doi.org/10.1080/00220485.2014.859955

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge Academic.

Crisp, G., Nora, A., & Taggart, A. (2009). Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An analysis of students attending a Hispanic serving institution. *American Educational Research Journal*, *46*(4), 924–942. https://doi.org/10.3102/0002831209349460

Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review*, *64*, 313–342. https://doi.org/10.1016/j.econedurev.2018.03.008

Deslauriers, L., Harris, S., Lane, E., & Wieman, C. (2012). Transforming the lowest-performing students: An intervention that worked. *Journal of College Science Teaching*, *41*, 80–88.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 763–782). Springer Science + Business Media.

Gordanier, J., Hauk, W., & Sankaran, C. (2019). Early intervention in college classes and improved student outcomes. *Economics of Education Review*, *72*, 23–29. https://doi.org/10.1016/j.econedurev.2019.05.003

Jimerson, S. R., Campos, E., & Greif, J. L. (2003). Toward an understanding of definitions and measures of school engagement and related terms. *California School Psychologist*, *8*, 7–27. https://doi.org/10.1007/BF03340893

Jones, B. D., Paretti, M. C., Hein, S. F., & Knott, T. W. (2010). An analysis of motivation constructs with first-year engineering students: Relationships among expectancies, values, achievement, and career plans. *Journal of Engineering Education*, *99*(4), 319–336. https://doi.org/10.1002/j.2168-9830.2010.tb01066.x

Kahu, E. R. (2013). Framing student engagement in higher education. *Studies in Higher Education*, *38*(5), 758–773. https://doi.org/10.1080/03075079.2011.598505

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lawton, S., & Taylor, L. (2020). Student perceptions of engagement in an introductory statistics course. *Journal of Statistics Education*, *28*(1), 45–55. https://doi.org/10.1080/10691898.2019.1704201

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal*, *37*, 153–184. https://doi.org/10.3102/00028312037001153

Moss, B. G., & Yeaton, W. H. (2015). Failed warnings: Evaluating the impact of academic probation warning letters on student achievement. *Evaluation Review*, *39*(5), 501–524. https://doi.org/10.1177/0193841X15610192

Muir, S., Tirlea, L., Elphinstone, B., & Huynh, M. (2020). Promoting classroom engagement through the use of an online student response system: A mixed methods analysis. *Journal of Statistics Education, 28*(1), 25–31. https://doi.org/10.1080/10691898.2020.1730733

Niranjan, S., Wu, J., & Jenner, C. (2015). Implications of student intervention and antecedents on academic motivation and success. *International Journal of Education Research, 10*(2), 1–21.

Ober, T. M., Hong, M. R., Rebouças-Ju, D., Carter, M. F., Liu, C., & Cheng, Y. (2021). Linking self-report and process data to performance as measured by different assessment types. *Computers & Education, 167*. https://doi.org/10.1016/j.compedu.2021.104188

Osborne, J., Parlier, R., & Adams, T. (2019). Assessing impact of academic interventions through student perceptions of academic success. *The Learning Assistance Review, 24*, 9–26. https://doi.org/10.1080/1360144X.2020.1777555

Pugatch, T., & Wilson, N. (2018). Nudging study habits: A field experiment on peer tutoring in higher education. *Economics of Education Review, 62*, 151–161. https://doi.org/10.1016/j.econedurev.2017.11.003

Rafi, Z., & Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: Replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology, 20*. https://doi.org/10.1186/s12874-020-01105-9

Schaufeli, W. B., Bakker, A. B., & Salanova, M. (2006). The measurement of work engagement with a short questionnaire: A cross-national study. *Educational and Psychological Measurement, 66*, 701–716. https://doi.org/10.1177/0013164405282471

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Smith, B. O., White, D. R., Kuzyk, P. C., & Tierney, J. E. (2018). Improved grade outcomes with an e-mailed "grade nudge". *The Journal of Economic Education, 49*(1), 1–7. https://doi.org/10.1080/00220485.2017.1397570

Tai, R. H., Liu, C. Q., Maltese, A. V., & Fan, X. (2006). Planning early for careers in science. *Science, 312*(5777), 1143–1144. https://doi.org/10.1126/science.1128690

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423–432. https://doi.org/10.1002/pits.20234

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement, 39*, 1–37. https://doi.org/10.1111/j.1745-3984.2002.tb01133.x

Whitney, B. M., Cheng, Y., Brodersen, A. S., & Hong, M. R. (2019). The scale of student engagement in statistics: Development and initial validation. *Journal of Psychoeducational Assessment, 37*(5), 553–565. https://doi.org/10.1177/0734282918769983

NOOSHIN KHOBZI ROTONDI
2000 Simcoe Street North
Oshawa, ON, L1G0C5