Journal of Social, Behavioral, and Health Sciences
2018, Volume 12, Issue 1, Pages 153–170
DOI:10.5590/JSBHS.2018.12.1.11

Security Mechanisms on Web-Based Exams in Introductory Statistics Community College Courses

Yelena Feinman
College of San Mateo

The credibility of unsupervised online exams is an ongoing concern in higher education. Proctoring, in the form of physical or remote supervision, has been the main mechanism for maintaining academic integrity. However, both forms of proctoring are expensive and inconvenient. Several researchers have examined security mechanisms as a substitute for proctoring and obtained mixed results. This article describes a quasi-experimental study, the main goal of which was to examine the effectiveness of nonbiometric security mechanisms. The security mechanisms were selected based on the taxonomy of cheating reduction techniques rooted in the fraud triangle theory. The security mechanisms were considered effective if the scores were equivalent or lower on the unproctored exams. Two one-sided dependent *t* tests were used to test for equivalence of scores on two sets of proctored and unproctored exams in face-to-face ($N = 704$), hybrid ($N = 91$), and online ($N = 55$) introductory statistics community college courses. In the first set, the proctored exam was followed by the unproctored exam; in the second set, the order was reversed. In the first set, the scores on proctored and unproctored exams were equivalent in face-to-face and online groups, but students in the hybrid group had significantly lower scores on the unproctored exam. In the second set, the students' scores were lower on the unproctored exam in all groups. The study's results suggest that the used security mechanisms were effective.

Keywords: *unproctored and proctored web-based exams, security mechanisms, taxonomy of cheating prevention techniques, fraud triangle theory, equivalence test*

Introduction

The latest development of information and communication technologies created favorable conditions for widespread adoption of learning management systems (LMSs) and integration of web-based assessment in everyday classrooms. By 2014, 99% of U.S. colleges and universities had at least one LMS that possessed a convenient and efficient way of delivering web-based exams (Dahlstrom, Brooks, & Bichsel, 2014). Over 85% of 170,000 faculty participating in the EDUCAUSE Core Data Service survey (Dahlstrom et al., 2014) responded that they used at least one available LMS for enhancing their teaching, including administering web-based assessments. The instructors valued the flexibility in creating online tests, customized feedback, immediate automatic recording of exam scores in the gradebook, and test items' analysis provided by the LMS. About 83% of 75,000 student-participants recognized the convenience of online tests and importance of immediate tests' feedback for their learning (Dahlstrom et al., 2014). At the same time, technological advances aggravated challenges associated with cheating, especially during unsupervised web-based exams (Shute &

This article is related to my dissertation "Alternative to Proctoring in Introductory Statistics Community College Courses." I would like to acknowledge my committee chair Dr. Deborah Y. Bauder and my methodologist Dr. Wade Smith for their valuable input and support during my dissertation journey. I also acknowledge my colleagues who assisted me in conducting this study.

Please address queries to: Yelena Feinman, Department of Mathematics, College of San Mateo, 1700 West Hillsdale Boulevard, San Mateo, CA 94402. Email: feinmanl@smccd.edu

Rahimi, 2017). One out of every four college students admitted to cheating with a smartphone during tests (Srikanth & Asmatulu, 2014); online collusion during asynchronous unproctored exams was detected among engineering students (de Sande, 2015), over 1,230 massive open online course students copied answers during asynchronous unproctored certificate exams using multiple online accounts (Northcutt, Ho, & Chuang, 2016), smartwatches were utilized for cheating with ease on an anatomy exam (Wong, Yang, Riecke, Cramer, & Neustaedter, 2017), and glasses with wireless cameras for transmitting exam questions were used to cheat by medical students (Parks, Lowry, Wigand, Agarwal, & Williams, 2018). The credibility of unsupervised assessment became the top challenge of online education (Instructional Technology Council, 2017). Proctoring is frequently used to maintain academic integrity (Lee-Post & Hapke, 2017). However, physical proctoring consumes time and money for both students and institutions and might be inconvenient or impossible for students who live far away from proctoring locations. Remote proctoring may not be suitable due to numerous technological requirements, high cost, and possible lack of effectiveness (C. Anderson & Gades, 2017). The disconnect between high demand in online testing and inability to maintain the credibility of unsupervised web-based exams without inconvenient and expensive proctoring constitutes a problem.

This article describes a quasi-experimental study, the main goal of which was to examine the effectiveness of nonbiometric security mechanisms by comparing students' scores on two sets of proctored and unproctored web-based exams. The criterion for the effectiveness of the mechanisms was equivalence of scores or lower scores on the unproctored exams. The scores were called equivalent if the difference between them was less or equal to 5% or 5 out of 100 points. The security mechanisms and structure of the web-based exams and their implementation are discussed. The article is intended for researchers, educators, administrators, policymakers, and other professionals working with web-based assessments. The experience of the implementation of the exams and study's results might be useful in any subject at any institution.

Numerous researchers have compared student performance during proctored and unproctored exams and obtained mixed results when using no, a few, or several security mechanisms. Sivula and Robson (2015) found that graduate students performed 34% better on an online unproctored exam without any security mechanisms. Similarly, Fask, Englander, and Wang (2015) did not use any security mechanisms and found that undergraduate students' performance on unproctored exams in an introductory statistics course was significantly better than on proctored exams. Arnold (2016), who used only two security mechanisms, randomization of multiple-choice questions and time restriction, found that first-year undergraduate students performed better on unproctored exams. Daffin and Jones (2018) found that without test time restrictions, psychology students had 20% higher scores on unproctored exams than on proctored exams. The findings of these researchers suggested that the use of no or a few security mechanisms results in significantly better student performance on unproctored exams.

Varble (2014), who incorporated randomization; restricted time; blocked backtracking, which does not allow going back to the previous question; and lockdown browser, which prevents accessing information from the Internet or computer, found that marketing university students did significantly better on unproctored web-based exams than on proctored pencil-and-paper test. The difference in scores was observed in all lower order thinking items. Ladyshewsky (2015) used the same security mechanisms as Varble, except lockdown browser, but incorporated higher order thinking questions. Ladyshewsky found that postgraduate business students performed better on the proctored pencil-and-paper exams than on unproctored web-based exams. The results of Varble and

Ladyshewsky suggest that lower order thinking questions may decrease the security of exams, while higher order thinking items may increase the security of exams.

Beck (2014) compared students' scores on secured proctored and unproctored exams in three sections of an economics course offered in face-to-face, hybrid, and online modes. In addition to randomization and time restriction, the researcher incorporated one question per page, blocked backtracking and cheating warning statement, but no lockdown browser. Beck's face-to-face and hybrid students took the proctored exam in pencil-and-paper format, while online students took the same exam in unproctored web-based format. The researcher found no significant difference in scores on proctored and unproctored exams in all course delivery modes. Similar to Beck, Stack (2015) used randomization, time restriction, one question per page, and blocked backtracking on unproctored exams in criminology courses. Stack incorporated lockdown browser, did not use a cheating warning statement, but administered the unproctored exams synchronously and found no significant difference in scores on pencil-and-paper proctored and web-based unproctored exams. Beck's and Stack's findings imply that the addition of cheating warning statement and synchronous testing may lead to no significant difference in students' scores on proctored and unproctored exams.

The previous researchers did not control for exam delivery mode when administering proctored exams in pencil-and-paper format and unproctored exams in a web-based format. This difference in test administration could influence students' scores (Bayazit & Aşkar, 2012; Jeong, 2014; Maguire, Smith, Brallier, & Palm, 2010). Beck (2014), Ladyshewsky (2015), and Stack (2015) did not ground the selection of the mechanisms in the fraud triangle theory or any other theory. Varble (2014) used the taxonomy for the selection of the mechanisms but did not discuss interactions between its components. The combination of the security mechanisms used in the present quasi-experiment had not been studied. Beck and Stack inferred comparability of scores on proctored and unproctored exams based on nonsignificant results. Beck studied the course delivery mode effect with small sample sizes in online ($N = 19$) and hybrid ($N = 20$) sections; Stack did not discuss whether the proctored and unproctored groups were comparable. None of the researchers considered the order in which proctored and unproctored exams were administered. The present study was conducted to address these gaps and investigate whether scores on proctored and unproctored automatically-graded web-based exams with the same security mechanisms were equivalent or lower on the unproctored exams. Additionally, the pattern of the scores with respect to the order of exams' administrations was examined.

The instructors involved in the study systematically selected security mechanisms to neutralize cheating during web-based exams. The selection of the security mechanisms was explained by the taxonomy of cheating reduction techniques (Varble, 2014) rooted in the fraud triangle theory (Cressey, 1950).

Fraud Triangle Theory

Cressey (1950) identified three major factors needed to commit fraud: opportunity, need, and rationalization. These factors were mapped onto an educational setting and used for understanding, prediction, and prevention of academic cheating. In education, asynchronous examinations and unlimited time on tests may increase the opportunity to cheat. The need to maintain a high grade point average (GPA) and be eligible for scholarships and prestigious universities may stimulate the need to cheat. Students usually rationalize their dishonest behavior by claiming that it is not clear what constitutes academic misconduct and no one gets caught (Tinkelman, 2012). The taxonomy derived from the theory has three categories: opportunity reduction, need reduction, and

rationalization reduction (Varble, 2014). The purpose of each category is to neutralize the corresponding cheating behavior generated by perceived opportunity, need, and rationalization. The opportunity reduction category may involve time restriction and higher order thinking level test items. The need reduction category emphasizes the true value of acquired knowledge and importance of the course content for a future profession. The rationalization reduction category may include institutional policies and cheating statements. Although the opportunity factor can be controlled by faculty the most, all three factors are important and can influence each other (Tinkelman, 2012). If any of the fraud triangle factors is reduced, neutralized, or blocked, less cheating should take place.

Security Mechanisms Used in the Study

To eliminate the opportunity for one student taking an exam at one time and then helping a classmate with the same exam at another time and prevent dissemination of exam items, the instructors used synchronous administration of the unproctored web-based exam. The students in all introductory statistics sections took each unproctored web-based exam on the same day during the same time frame. Although synchronous testing is considered one of the strongest security mechanisms (de Sande, 2015; Northcutt et al., 2016), students may have schedule conflicts with it. To reduce possible schedule conflicts, the dates and times of all unproctored exams were announced and posted on the course web-site on the first day of classes. To neutralize collusion when two or more students work on the same exam side by side, the instructors used restricted time, randomization, one question per page, and blocked backtracking. The test time was carefully identified such that the allocated number of minutes was sufficient to complete the questions but not sufficient to look up the answers on the Internet, in printed sources, and call or text friends. With randomization of test items, students sitting next to each other saw different questions of the exam. One question per page and blocked backtracking eliminated the opportunity to go back and insert the answer found by another person. To neutralize the opportunity to find solutions on the Internet or in printed sources, the instructors incorporated higher order thinking test items. Answering these questions required statistical reasoning, critical thinking, and interpretation, which reflected the main focus of the inquiry-based curriculum used by the faculty. To prevent the distribution of answers while the exam is still open, the instructors used deferred feedback: Examinees did not know whether their answers were correct. Multiple versions of the same web-based exam for students who could not take the test at the designated time and making the exams inaccessible right after the tests' submissions further decreased circulations of exam items among the students.

To reduce the need to cheat generated by fear of getting bad grades and rationalization that the test was too hard, students were given a web-based practice test before each exam, the structure and time frame of which were identical to the actual exams. Additionally, all needed formulas were provided on each exam. Discussions about the departmental focus on credibility of the offered courses and high standards eliminated rationalization that the use of security mechanisms is unfair. To prevent rationalization of not knowing what constitutes cheating, the instructors developed a common syllabus with clearly stated cheating policies and consequences of academic dishonesty. Before each unproctored exam, the cheating warning statement was emailed to students and posted on the course website. To neutralize the need to cheat, the instructors built a high-quality teaching and learning environment and an atmosphere of mutual respect, emphasizing the true value of education.

Lockdown browser was not used in the study because of several reasons. This security mechanism was not available at the college where the study took place. Lockdown browser may not be as effective in preventing the use of the Internet, emailing, and copying test items during exams as it

was before due to the high popularity of mobile technology. By 2015, about 97% of college students had portable devices that they carried around on a regular basis (Walters & Hunsicker-Walburn, 2015). Access to the Internet, blocking of which is the main purpose of a lockdown browser, may become available on these devices instantaneously just with one click (Walters & Hunsicker-Walburn, 2015). Moreover, each opportunity prevention technique should be used when it is highly needed because it may increase rationalization and trigger more cheating (Walters & Hunsicker-Walburn, 2015). The Internet is not very useful when higher order thinking exam questions uniquely created by faculty are incorporated (Ladyshewsky, 2015). Even if students go on the Internet during exams, they are not able to find the answers there.

Method

A quasi-experimental one-group sequential design was used with archived scores of introductory statistics students on two sets of secured proctored and unproctored web-based exams administered from Fall 2015 through Summer 2017. This within-subject design allowed for controlling for initial differences among the participants. The proctored exam format was considered the control condition, while the unproctored format was considered the experimental one. Each student went through both conditions by taking two sets of proctored and unproctored exams in a certain sequence. In Set 1, which took place in the middle of each semester, the proctored web-based exam was followed by the unproctored one. In Set 2, which was administered at the end of each term, the order was reversed. The retest interval within each set was 7–10 days; the retest interval between the sets was 1 month. The test–retest intervals of 7–10 days and 1 month were dictated by the course curriculum.

The faculty decided to administer the first web-based exam in a proctored format, assuming that students would feel more comfortable to complete a new type of assessment in a classroom environment; the alternative form of the first exam was administered in an unproctored format. The alternative form of the exam had the same items but with different numerical values and themes. Because the instructors wanted to use in-class time at the end of the semester for preparation for the final exam, the first web-exam in Set 2 was administered in an unproctored environment, while the alternative form of this exam was proctored. This sequence of the exams occurring in a natural educational setting allowed for examining the pattern of scores with respect to the order in which proctored and unproctored exams were administered and interpreting possible fatigue, practice, and learning effects. Additionally, the study design controlled for other variables that could influence the relationship between the exam format and students' scores. All involved faculty used the same materials, curriculum, assessments, and syllabus, which minimized instructor effect. All proctored exams took place on the same day in the same classroom and were proctored by the instructors, which allowed controlling for history and proctored environment effects. The unproctored exams could be completed at any location with Internet accesses. However, the students were advised to take the unproctored exams in a quiet environment free of any distraction. All exams were automatically scored by an LMS, Moodle, which reduced grading effect.

Setting and Sample

The study's setting was a suburban community college, which serves 9,000 students every semester. About 82% (n = 9 out of 11) of the college transfer programs have introductory statistics as a requirement. The students' scores in all web-based introductory statistics sections offered by the college were analyzed in the study. A total of 850 students took at least one study's exam: 57% females and 43% males. The participants' ages ranged from 14 to 50 years, with the mean of 22; the mean GPA was 3.19. The GPAs were requested from the institutional research department.

Out of the 850 participants, 704 were face-to-face students, 91 hybrid, and 55 online. While most students took both exams in each set with all security mechanisms, there were students who could not take the unproctored exams at the scheduled time and took the alternative version of the same exam at different time, students who had extended test time, and students who took the second exam in Set 2 in asynchronous unproctored format. In each group, there were students who did not take one or both exams in Set 1, but took one or both exams in Set 2. The study's design and the number of students on each exam are shown in Figure 1.

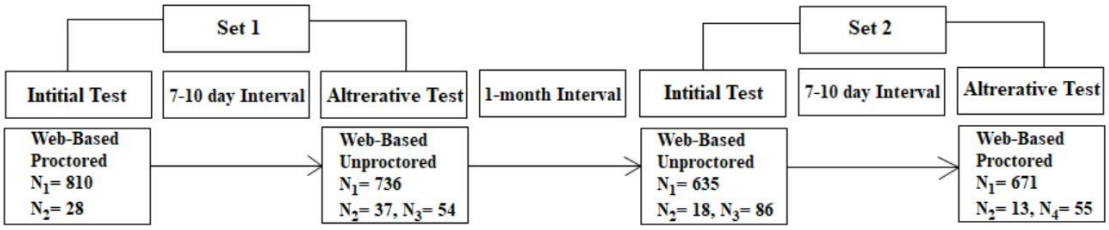

Figure 1. *The Study's Design With the Number of Students on Each Exam. N1 = number of students who took the exam with all security mechanisms; N2 = number of students who had extended time; N3 = number of students who had schedule conflict; N4 = number of students who took the second exam in Set 2 in unproctored format.*

Instrumentation

Each study's exam was created by faculty in accordance with high standards and had three drop-down, four multiple-choice, and 16 short-answer questions, all of which were automatically scored by the LMS Moodle 3.0. All exams involved in the study were a part of regular educational practice. The choice of the number of questions was made based on the number of concepts covered by the exams, the time instructors could allocate for in-class proctored tests, and recommendations found in the literature. By piloting administration of the exams in a proctored environment, it was identified that 70 min for 23 questions is sufficient to complete the assessment without rushing. Ladyshewsky (2015) had similar recommendations with respect to allocated test time.

Multiple-choice and drop-down questions were selected to measure statistical reasoning and interpretation. For questions, answers to which require calculation, the instructors used short-answer format. To reduce opportunities to guess, the number of multiple-choice and drop-down items was minimized. The questions in all exams were scrutinized for quality; there were no cued or overlapping items. Each exam question was designed to be answered independently from others such that randomization and blocked backtracking could not impact student performance. To align each exam item with needed cognitive processes and knowledge dimensions, the instructors used the revised Bloom's taxonomy (L. W. Anderson, Krathwohl, & Bloom, 2001) and more detailed taxonomy classification done by Darwazeh and Branch (2015).

The exams within each set were alternative; the exams between the sets were on different but equivalent topics. All four exams had the same cognitive and conceptual levels of difficulty; the same

structure; the same number of multiple-choice, drop-down, and short-answer questions; the same allocated time; and the same security mechanisms. The same number of points was assigned to the corresponding questions. The same exams were used in all sections during the study's time frame. The samples of exams' questions are provided in Figure 2.

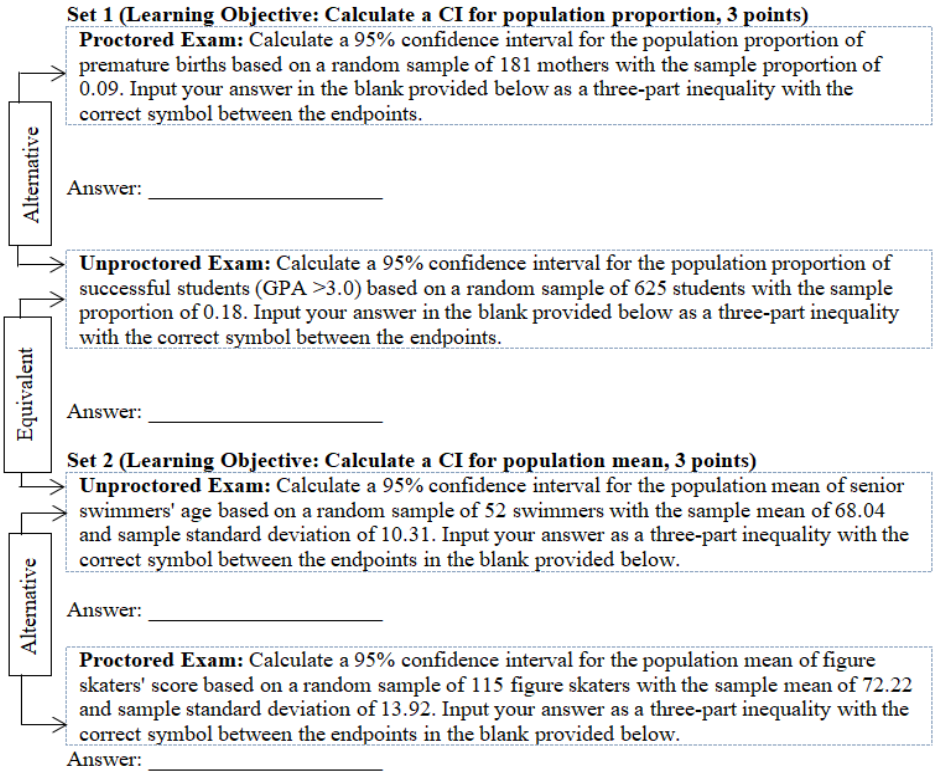

Figure 2. *Sample Exam Questions. CI = confidence interval; GPA = grade point average.*

Data Collection and Analyses

After institutional review board approval (05-23-17-0315459) was obtained, the institution where the study took place provided the exam scores and demographics of all students enrolled in web-based statistics sections offered in face-to-face, hybrid, and online formats from the Fall 2015 through Summer 2017 semesters. The data screening indicated no missing entries but revealed that some students took the second exam in Set 2 in asynchronous unproctored format, instead of proctored format. Because archived data collected in a natural educational setting were used, no actual recruitment took place.

To test for equivalence of scores on proctored and unproctored exams, two one-sided dependent *t* tests (TOSTs) with the corresponding 90% confidence interval (CI) were used (Lakens, 2017a; Rogers, Howard, & Vessey, 1993; Schuirmann, 1987; Seaman & Serlin, 1998). As recommended by Lakens, Scheel, and Isager (2018), to improve interpretations of the results of TOST, a null-hypothesis significance tests (NHST) with the corresponding 95% CI were applied as well. Because

TOST has not been commonly used in educational research (Briones & Benham, 2017), this procedure will be briefly described.

In TOST, the null (H_0) and alternative (H_a) hypotheses can be represented in the following form:

$$H_0: M_1 - M_2 \geq \Delta \text{ and } M_1 - M_2 \leq \Delta$$

$$H_a: -\Delta < M_1 - M_2 < \Delta,$$

where an apriori equivalence bound Δ is chosen based on the criterion of how far the two means can differ while being considered equivalent (Lakens, 2017a). Thus, the null hypothesis consists of two one-sided *t* tests, and to establish statistical equivalence, both one-sided *t* tests should be statistically significant. The two *t* statistics t_{upper} and t_{lower} with their corresponding *p* values p_{upper} and p_{lower} are calculated; the larger of the two ps with its corresponding *t* statistics are reported (Lakens, 2017a). The equivalence bound ideally should be determined in raw mean difference scores based on practical consideration, theoretical predictions, or prior research (Lakens et al., 2018). In equivalence testing, a CI with the confidence level of 90% is calculated because this CI corresponds to two one-sided tests with $\alpha = 0.05$: $(1 - 2 \times \alpha) \times 100\% = 90\%$. A 90 % CI around the observed mean difference, in addition to 95% NHST CI, adds graphical and numeric representation to the TOST and NHST results (Lakens, 2017a). If 90% TOST CI is not entirely inside of [–Δ, Δ] and NHST CI excludes 0, the means are not equivalent and statistically different. When 90% CI is entirely inside of [–Δ, Δ] and 95% NHST CI includes 0, the means are equivalent and not different. If 90% CI is entirely inside of [–Δ, Δ] and 95% NHST CI excludes 0, the means are statistically different, but the difference is small and the means are considered equivalent. When 90% CI is not entirely inside of [–Δ, Δ] and 95% NHST includes 0, statistical indeterminacy occurs: The means are not statistically different and not statistically equivalent (Lakens, 2017a).

In the given investigation, an equivalence bound of 5 points out of 100 (5%) was chosen based on the practical considerations that exam grades were set at every 5% increase (A = 95–100%, A– = 90–95%, etc.) and previous studies with 100-point scale exams (Bogacki, Best, & Abbey, 2004; Rusticus & Lovato, 2011, 2014). An Excel spreadsheet developed by Lakens (2017b) was used for the TOST, NHST, 90% CI and 95% NHST CI analyses, and for calculation of the effect sizes. Software developed by Uanhoro (2017) was used to find CIs for effect sizes. Graphical representations of the 95% and 90% CIs were obtained through R script created by Lakens (2017c).

Results

All results are reported as significant at $\alpha = .05$. The Hedges' g_{av}, Hedges' g_s, and common language effect sizes and their confidence intervals are included. Although *t*-test statistics and *p* values for both TOSTs were found, only the larger *p* and corresponding *t* are reported. A priori power analysis for the dependent *t* test with the alpha of .05, medium effect size of 0.5, and power level of .80 indicated the sample size of 17 per group.

Descriptive Statistics

The test scores of the students who could not take the unproctored exams at the scheduled time were excluded from the statistical analysis because not all security mechanisms were used by these students. The scores of the students with extended test time were not included in the analysis as well due to the small sample sizes of 16 in Set 1 and 13 in Set 2. For all other students, the comparison of their individual scores on the proctored and unproctored exams was done within each

course delivery mode first in Set 1 and then in Set 2. The descriptive statistics of these scores are provided in Table 1.

Table 1. *Descriptive Statistics of Scores Within Course Delivery Modes in Set 1 and Set 2*

	Set 1				Set 2			
Group	*N*	M_P (*SD*)	M_U (*SD*)	Diff	*N*	M_P (*SD*)	M_U (*SD*)	Diff
Face to face	599	66.17	68.30	–2.13	469	63.69	69.46	–5.77
		(21.16)	(21.10)	(24.31)		(19.43)	(17.05)	(13.82)
Hybrid	85	71.35	66.93	4.42	78	64.87	69.80	–4.93
		(20.53)	(22.27)	(11.84)		(19.33)	(20.53)	(13.04)
Online	48	73.59	74.68	–1.09	46	64.40	69.56	–5.16
		(20.94)	(18.57)	(13.79)		(25.00)	(20.41)	(14.59)

Note. P = proctored; U = unproctored; Diff = P – U.

As seen in Table 1, in Set 1, the score means were just slightly higher on the unproctored exam than on the proctored one in the face-to-face and online groups. In the hybrid group, the score means were higher on the proctored exam. In Set 2, the scores were lower on the unproctored exam in all groups. Out of all four exams, regardless of the group, the smallest mean scores were earned on the first exams in Set 2. In both sets, in all groups, except the hybrid group in Set 1, the students performed better on the exam that was administered second. The inspection of the individual data points revealed that about 67%, 86%, and 71% of students in the face-to-face, hybrid, and online groups in Set 1 had a score difference less than 5% or performed better on the proctored exams. In Set 2, these values were 79%, 87%, 80%, respectively. In the hybrid group, the proportion of students with equivalent scores or higher scores on the proctored exams was almost the same in both sets. In the face-to-face and online groups, the pattern of scores was similar within each set.

Results of Testing Assumptions for Dependent *t* Test

The dependent *t* test yields trustworthy results if participants are randomly sampled from the population for which inferences are made. Additionally, the difference scores should be independent of each other and normally distributed (Lakens, 2017a). The participants were not randomly selected from the population of all community college students who take introductory statistics in web-assisted environments, which is discussed in the Limitations section of this article. The difference scores were independent of each other. The kurtosis, skewness, and the Shapiro–Wilk test results for the difference scores in each group are shown in Table 2.

Table 2. *Results of Testing Assumptions for the Difference Scores*

Group	Skewness	Kurtosis	Shapiro–Wilk test
Set 1			
Face to face	0.16	1.06	*
Hybrid	1.26	4.43	*
Online	–2.02	4.20	*
Set 2			
Face to face	0.10	0.20	.53
Hybrid	0.85	4.43	*
Online	–0.53	0.37	.23

* $p < .001$.

As seen in Table 2, the Shapiro–Wilk test was significant in all groups except the face-to-face and online groups in Set 2. However, the *t*-test procedure is robust to nonnormality with sample sizes bigger than 30, as it was in all groups.

Results of TOST and NHST Analyses

The results of TOST and NHST with the corresponding CIs for the difference scores in all subgroups are summarized in Table 3. Hedges' g_{av} effect size with its CI and common language effect size are also provided in the table.

Table 3. *Summary of Two One-Sided Dependent t Test (TOST) and Null-Hypothesis Significance Test (NHST) Findings*

	TOST			NHST					
Subgroup	*t*	*p*	[90% CI]	*t*	*p*	[95% CI]	g_{av}	[95% CIg_{av}]	ES_{CL}
Set 1									
Face to face[a,b]	2.9	<.001	[–3.8, –0.5]	–2.1	<.001	[–4.1, –0.2]	0.1	[0, 0.2]	.5
Hybrid[b]	–0.4	.3	[2.3, 6.6]	3.4	<.001	[1.9, 7.0]	0.2	[0.1, 0.3]	.6
Online[a]	1.9	<.001	[–4.5, 2.2]	–0.6	.6	[–5.2, 2.8]	0.1	[–0.1, 0.3]	.5
Set 2									
Face to face[b]	–1.4	.9	[–7.0, –4.8]	–9.2	<.001	[–7.1, –4.6]	0.3	[0.2, 0.4]	.6
Hybrid[b]	0.1	.5	[–7.1, –2.7]	–3.7	<.001	[–7.6, –2.3]	0.2	[0.1, 0.4]	.6
Online[b]	–0.1	.5	[–8.8, –1.6]	–2.4	<.001	[–9.4, –0.8]	0.2	[0, 0.4]	.6

Note. CI = confidence interval; CL = common language.
[a] Statistically equivalent at $\Delta = 5$. [b] Statistically different.

As seen in Table 3, in the Set 1 face-to-face group, on average, the scores were statistically equivalent and statistically different, indicating that the found difference was not large enough for the scores to be not equivalent at $\Delta = 5\%$. In the hybrid group, the scores were not equivalent and significantly lower on the unproctored exam. In the online group, the scores on the proctored and unproctored exams were statistically equivalent. In Set 2, the scores were not equivalent and significantly lower on the unproctored exams in all groups.

Results of Additional Tests

The data screening revealed that 55 students completed second exam in Set 2 in unproctored format, instead of planned proctored; 51 of these 55 students completed both exams in Set 2. These 51 students took the first exam synchronously and the second exam asynchronously. The dependent *t* test and TOST were used to compare the scores on these two unproctored exams. The scores on the asynchronous unproctored exam (M_{asynch} =68.88, SD_{asynch} =20.39) were not equivalent, $t(50) = –1.93$, $p = 0.97$, 90% CI [–13.99, –5.62], and 9.8% significantly higher than the scores on the synchronous unproctored exams (M_{synch} = 59.07, SD_{synch} = 18.62), $t(50) = –3.93$, $p < .001$, 95% CI [–14.83, –4.79], Hedges' $g_{av} = 0.49$, 95% CI for g_{av} [0.23, 0.77], $ES_{CL} = .71$).

The students in face-to-face, hybrid, and online sections were compared with respect to GPA and age. The descriptive statistics of the variables GPA and age are provided in Table 4.

Table 4. *Descriptive Statistics of Grade Point Average (GPA) and Age*

Group	*N*	M_{GPA}	SD_{GPA}	M_{Age}	SD_{Age}
Face to face	704	3.15	0.53	21.59	4.35
Hybrid	91	3.13	0.58	25.29	6.80
Online	55	3.23	0.55	21.98	4.57

As seen in Table 4, the GPA was similar across the course delivery modes while students in the hybrid sections were about 3.5 years older than students in face-to-face and online sections. To test whether the observed similarities in GPA and difference in age were significant, the statistical tests were conducted.

A power analysis for the independent t-test with the alpha of .05, medium effect size of .5, and power of .80 determined the sample size of 53 per group. The Levene's test was not significant for GPA (p = .344), but was significant for age (p < .001). The sample sizes were not equal. As recommended by Lakens (2017a), to control for unequal sample sizes and unequal variances, TOST and NHST Welch's t tests were used. According to Armstrong (2014), the Bonferroni correction for conducting multiple t tests is not advised if only a few planned comparisons are incorporated, as it is in the GPA and age analyses. For this reason, corrections for multiple comparisons were not applied. Based on the discussions with relevant stakeholders, the equivalence bound $\Delta = 0.3$ points and $\Delta = 2$ years were used for GPA and age, respectively.

The results of the tests and the corresponding effect sizes with their CIs are summarized in Table 5.

Table 5. *Grade Point Average (GPA) and Age Comparison Across the Course Delivery Modes*

	TOST			NHST					
Group	t	p	[90% CI]	t	p	[95% CI]	g_s	[95% CIg_s]	ES_{CL}
GPA									
Face to face vs. hybrid[a]	31.0	<.001	[–0.1, 0.1]	0.3	.2	[–0.2, 0.2]	0.2	[0.1, 0.3]	.6
Face to face vs online[a]	2.8	<.001	[–0.2, 0.0]	–1.1	.3	[–0.3, 0.1]	0.2	[–0.1, 0.4]	.6
Hybrid vs online[a]	–1.4	.2	[–0.3, 0.1]	–1.4	<.01	[–0.4, 0.2]	0.2	[–0.2, 0.5]	.6
Age									
Face to face vs hybrid[b]	-3.3	.9	[-4.9, -2.5]	-5.1	<.001	[-5.1, -2.3]	0.8	[.6, 1.01]	.7
Face to face vs online[c]	2.5	<.001	[-1.5, 0.7]	-0.6	.5	[-1.7,0.9]	0.1	[-.4, .2]	.5
Hybrid vs online[b]	1.4	.9	[1.8, 4.9]	3.5	<.01	[1.5,5.2]	0.5	[.2, .9]	.7

Note. CI = confidence interval; CL = common language.
[a] Statistically equivalent at $\Delta = 0.3$. [b] Statistically different. [c] Statistically equivalent at $\Delta = 2$.

The GPAs of face-to-face, hybrid, and online students were equivalent at $\Delta = 0.3$ and not statistically different. The hybrid students were significantly older than face-to-face students and online students. The ages of the face-to-face and online groups were equivalent at $\Delta = 2$ and not statistically different.

Reliability and factor analyses, the results of which are provided in Table 6, were conducted for all four exams. As seen in Table 6, the number of students who responded to all test items was bigger on the proctored exams than on the unproctored exams. The model fit indices were similar across all four exams; the reliability and construct validity were adequate (all $\alpha \geq .79$; all $df < 2$; all goodness-of-

fit index, adjusted goodness-of-fit index, and comparative fit index > .90; all root mean square residual and root mean square error of approximation < .05).

Table 6. *Results of Reliability and Confirmatory Factor Analysis on All Four Exams*

Group	*N*	α	χ^2	χ^2/df	GFI	AGFI	CFI	RMR	RMSEA
Set 1									
Proctored	536	.86	401.49	1.75	.94	.93	.92	.03	.03
Unproctored	439	.79	339.40	1.48	.94	.92	.90	.03	.03
Set 2									
Proctored	444	.86	402.88	1.76	.93	.91	.91	.04	.04
Unproctored	653	.79	415.70	1.83	.95	.93	.90	.04	.04

Note. GFI = goodness-of-fit index; AGFI = adjusted GFI; CFI = comparative fit index; RMR = root mean square residual; RMSEA = root mean square error of approximation.

Discussion

In all groups, none of the study's statistical tests revealed significantly higher scores on unproctored exams. Thus, on average, the students' scores were either equivalent or lower on the unproctored exams. The same pattern was observed at the individual level: The majority of the students, regardless of the course delivery mode, had the score difference less or equal to 5% or performed better on the proctored exams. These findings suggest that the combination of the security mechanisms was effective: If the students attempted to cheat during the unproctored exams, they were unsuccessful.

In Set 1, in the face-to-face and online groups, the scores were 1.1% and 2.1% higher on the unproctored exam than on the proctored exam. These differences were small enough for the score to be statistically equivalent at $\Delta = 5\%$. In the hybrid group, the scores were about 4.4% lower on the unproctored exam than on the proctored exam, and this difference was significant for the scores not to be equivalent at $\Delta = 5\%$. The different results in the hybrid group cannot be attributed to distinct academic abilities because GPA was equivalent between all course delivery modes. A more suitable explanation can be related to the significantly higher age of hybrid students than face-to-face and online students. Ladyshewsky (2015) observed that older hybrid students tended to have lower scores on unproctored exams than younger face-to-face students and explained it by a possible higher level of business of older adults resulting in more distractions at home. The relationship between older age, lower scores on unproctored exams, and distractions in an unproctored environment may be investigated in future studies. The distinct pattern of scores in the hybrid group contradicts Beck's (2014) results, who did not find a significant difference in performance across the course delivery modes. Unlike the present study, Beck's hybrid students were not significantly different in age from face-to-face and online students. To increase the generalizability of the results in Set 1, replication of the study in other institution and different populations of students is recommended.

In Set 2, in all groups, the scores on the unproctored exam, which was administered first, were between 4.9% and 5.9% significantly lower than the scores on the proctored exams. This raises the question of why, unlike Set 1, the students' performance was lower on the unproctored exam. Because in all groups the scores on the first exam in Set 2 were the lowest out of all four exams, it is suitable to assume that the students were not prepared for this exam as well as they were prepared for other exams. The low performance could be explained by an end-of- the-semester fatigue effect, which the students were able to overcome on the last exam. Forgetting of some previous knowledge is another suitable explanation. The first exam in Set 2 took place in 30 days after Set 1 exams. In 30

days, individuals can forget up to 90% of acquired information and skills (Falleti, Maruff, Collie, & Darby, 2006). To test whether the decrease in scores could be explained by end-of-the-semester fatigue and forgetting, the study can be replicated with the reduced retest interval between the sets from 30 to 7–10 days.

In both sets in all groups, except the hybrid group in Set 1, the students performed better on the second exam regardless of the order in which the proctored and unproctored exams were administered. It can be said that a practice or retake effect took place. However, reduction in a practice effect occurs when alternative forms (Benedict & Zgaljardic, 1998) and randomization of test items (Falleti et al., 2006) are used because individuals perceive a retest as a new exam. Additionally, Randall and Villado (2016) found that the use of security mechanisms that minimize opportunities to copy and disseminate exam questions diminish score contamination due to retesting. In the present study, alternative forms and the security mechanisms were used. Moreover, the students did not know that the second exam in each set was an alternative version of the first one. Thus, most likely, the students perceived each test as a new exam. The practice tests administered since the beginning of each semester could eliminate retest score increase due to becoming familiar with the test form and structure. A more suitable explanation of the higher scores on the second exam in each set is a learning effect. In both sets, the students studied for the first exam, took the exam, understood what concepts were not mastered and studied again, which resulted in higher scores on the second exams.

In the group of students who took the first exam in Set 2 in unproctored synchronous format and the second exam in unproctored asynchronous format, the scores were 9.8% significantly higher on the asynchronous exam than on the synchronous unproctored exam. This result can be explained by cheating, which reinforces the utilization of the synchronous testing. For comparison, the difference in scores in Set 2 of the students who took the second exam in the proctored format was about 5%. The relationship between asynchronous and synchronous administration of unproctored exams and students' scores can be studied further in future research.

The finding that the number of students who responded to all test items was larger on the proctored exams than on the unproctored exams may indicate that in unproctored environments students tend to skip more questions than in proctored environments. Searching for answers on the Internet or through other sources while taking the test and not having enough time to respond to all questions is one possible explanation. However, it does not explain why students miss not only the last questions of the assessment, but also items in the middle of the test. The relationship between the exam format, proctored versus unproctored, and the number of students who respond to all test items may need to be studied in future research.

Faculty and Students' View on the Security Mechanisms and Exams

At the department where the study took place, the incorporation of the security mechanisms and implementation of web-based exams was faculty-driven. The instructors valued the many advantages associated with web-based testing: convenience, flexibility, automatic grading, immediate test item analysis, and opportunity to use more in-class time on instruction and learning by administering some exams at home. The faculty selected the security mechanisms fully understanding why each of the mechanisms was needed and important. The implementation of the exams went smoothly, except one obstacle: several students, predominantly from face-to-face sections, had schedule conflicts and could not take unproctored exams synchronously. The instructors overcame this obstacle by creating alternative versions of the unproctored exams and

administering them after the scheduled exams. In the discussions with each other, the faculty shared that the use of the security mechanisms added confidence in administering unsupervised exams. The instructors explained to their students the purpose of the web-based exams and security mechanisms. The faculty's impression was that many students liked the convenience of web-based exams and understood the necessity of the use of the security mechanisms. The instructors have begun administering secured web-based tests in other math courses and using online exams for student learning outcomes assessment at the departmental level. An individual instructor's opinion about each security mechanism can be investigated in a follow-up study.

The level of exam security and choice of mechanisms may depend on institutional policies on academic misconduct, educational standards, accreditation requirements, and articulation with other colleges and universities. Maturity and culture of the student population, the number of course sections taking the same exam, and goals and type of assessment can be considered as well. At the college where the present study took place, the institutional policies on academic dishonesty are required to be included in a syllabus and clearly explained to students. The department focuses on high academic standards and requires the administration of summative exams in all math courses. Accreditation requirements include the use of security on unsupervised summative exams. Articulation with colleges and universities requires keeping the same academic standards of a course across all modes in which the course is delivered.

In the present study, up to 11 sections of introductory statistics completed the same exams each semester. These community college students took many previous courses together and knew each other well, which created favorable conditions for sharing information about exams questions across the sections. For these reasons, the higher level of security on unsupervised summative exams was needed. The practice exams were used as a learning tool rather than a testing tool. The faculty did not use synchronous testing or deferred feedback on the practice exams.

The combination of the security mechanisms used by the department can be tailored to the needs of a particular institution and instructor. Thus, to improve the credibility of unsupervised web-based exams, some, most, or all of the study's security mechanisms can be used by faculty at other institutions.

Limitations

The major limitation of the study was its quasi-experimental nature. The sample of students in the study was not randomly selected from the population of all community college students who take introductory statistics in web-assisted environments. The participants were not randomly assigned to proctored and unproctored exams or their classes with respect to the course delivery modes because the study took place in a natural educational setting. A great deal of effort was put to minimize this limitation by selecting a design that controls for initial differences in subjects under investigation and in any variable that could potentially influence the relationship between the exam format and students' scores.

Although the study's sample was not random, it represented the population well. Introductory statistics, a traditional four-unit transferable course offered by all community colleges in the state, is required for most transfer majors. Therefore, the sample was heterogeneous with respect to majors and included typical community college students. Additionally, the web-based exams were administered in a natural educational setting as a part of regular educational practices. The findings

of high-quality quasi-experiments conducted in natural educational settings might be more applicable than findings of true randomized experiments because randomization is almost never possible in a regular educational practice (Kim & Steiner, 2016). Thus, the results of the study can be generalizable for similar institutions with a similar population of students.

How the Study Advances the Previous Research

To the best of my knowledge, this is the first study to investigate the effectiveness of security mechanisms for community college students. It is the first study on this topic in which both proctored and unproctored exams were delivered in a web-based form. The study adds to the previous research by incorporating not only opportunity-reduction techniques, but also the need- and rationalization-reduction techniques, considering the interaction between all three factors.

Conclusion

The era of classroom web-based assessment has begun. Proctored and unproctored web-based exams are in high demand among students and instructors. The present study's results suggest that, with the combination of security mechanisms used in the investigation, the credibility of unproctored exams might be comparable with the credibility of proctored exams. The findings empirically verify that the taxonomy of cheating prevention techniques rooted in the fraud triangle theory is an adequate theoretical framework for identifying effective security mechanisms. This theoretical framework can be used to secure web-based exams in any subject at any institution. The instructors' experience with the web-based exams' implementation adds to the body of the best practices of secured online assessment. The use of the security mechanisms utilized in the study may allow for assessing student knowledge in a credible, inexpensive, and convenient way, not spending valuable in-class time on testing in face-to-face and hybrid classes and enhancing viability of online courses. More students with full-time jobs and family commitments will be able to take online exams and obtain credible degrees.

References

Anderson, C., & Gades, P. (2017, June). *Proctoring exams in an online environment*. Paper presented at Innovate! Teaching With Technology, Morris, MN.

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Boston, MA: Allyn & Bacon.

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*, 502–508.

Arnold, I. J. (2016). Cheating at online formative tests: Does it pay off? *The Internet and Higher Education, 29*, 98–106. doi:10.1016/j.iheduc.2016.02.001

Bayazit, A., & Aşkar, P. (2012). Performance and duration differences between online and paper-pencil tests. *Asia Pacific Education Review, 13*, 219–226. doi:10.1007/s12564-011-9190-9

Beck, V. (2014). Testing a model to predict online cheating -much ado about nothing. *Active Learning in Higher Education, 15*, 65-75. doi:10.1177/1469787413514646

Benedict, R. H., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology, 20*, 339–352. doi:10.1076/jcen.20.3.339.822

Bogacki, R. E., Best, A., & Abbey, L. M. (2004). Equivalence study of a dental anatomy computer-assisted learning program. *Journal of dental education, 68,* 867–871.

Briones, E. M., & Benham, G. (2017). An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student samples. *Behavior research methods, 49*, 320–334.

Cressey, D. R. (1950). The criminal violation of financial trust. *American Sociological Review, 1*, 738–743. doi:10.2307/2086606

Dahlstrom, E., Brooks, C., & Bichsel, J. (2014). *The current ecosystem of learning management systems in higher education: Student, faculty, and IT perspectives*. Louisville, CO: ECAR. Retrieved from https://library.educause.edu/~/media/files/library/2014/9/ers1414-pdf.pdf

Daffin Jr., L. W., & Jones, A. A. (2018). Comparing student performance on proctored and non-proctored exams in online psychology courses. *Online Learning, 22,* 1.

Darwazeh, A., & Branch, R. (2015, November). *A revision to the* revised *Bloom's taxonomy*. Retrieved from https://members.aect.org/pdf/Proceedings/proceedings15/2015i/15_04.pdf

de Sande, J. C. G. (2015). Calculated questions and e-cheating: A case study. *Education Applications & Developments Advances in Education and Educational Trends Series, 2*, 91–99.

Falleti, M. G., Maruff, P., Collie, A., & Darby, D. G. (2006). Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. *Journal of Clinical and Experimental Neuropsychology, 28*, 1095–1112. doi:10.1080/13803390500205718

Fask, A., Englander, F., & Wang, Z. (2015). On the integrity of online testing for introductory statistics courses: A latent variable approach. *Practical Assessment, Research, & Evaluation, 20*, 1–12.

Instructional Technology Council. (2017). *ITC annual national eLearning report: 2016 Survey results*. Retrieved from https://associationdatabase.com/aws/ITCN/asset_manager/get_file/154447?ver=275

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology, 33*, 410–422. doi:10.1080/0144929X.2012.710647

Kim, Y., & Steiner, P. (2016). Quasi-experimental designs for causal inference. *Educational Psychologist, 51*, 395–405. doi:10.1080/00461520.2016.1207177

Ladyshewsky, R. K. (2015). Post-graduate student performance in "supervised in-class" vs. "unsupervised online" multiple choice tests: Implications for cheating and test security. *Assessment & Evaluation in Higher Education, 40*, 883–897. doi:10.1080/02602938.2014.956683

Lakens, D. (2017a). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*, 355–362. doi:10.1177/1948550617697177

Lakens, D. (2017b). *Spreadsheet to perform TOST equivalence tests*. Retrieved from https://osf.io/q253c/

Lakens, D. (2017c). *TOST in R*. Retrieved from https://cran.r-project.org/web/packages/TOSTER/index.html

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*, 259–269. doi:10.1177/2515245918770963

Lee-Post, A., & Hapke, H. (2017). Online learning integrity approaches: Current practices and future solutions. *Online Learning, 21*, 1–11. doi:10.24059/olj.v21i1.843

Maguire, K. A., Smith, D. A., Brallier, S. A., & Palm, L. J. (2010). Computer-based testing: A comparison of computer-based and paper-and-pencil assessment. *Academy of Educational Leadership Journal, 14*, 117.

Northcutt, C. G., Ho, A. D., & Chuang, I. L. (2016). Detecting and preventing "multiple-account" cheating in massive open online courses. *Computers & Education, 100*, 71–80.

Parks, R. F., Lowry, P. B., Wigand, R., Agarwal, N., & Williams, T. L. (2018). *Why students engage in cyber-cheating through a collective movement: A case of deviance and collusion*. Retrieved from https://www.researchgate.net/profile/Paul_Lowry/publication/324525509

Randall, J. G., & Villado, A. J. (2016). Take two: Sources and deterrents of score change in employment retesting. *Human Resource Management Review, 27*, 536–553. doi:10.1016/j.hrmr.2016.10.002

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*, 553.

Rusticus, S. A., & Lovato, C. Y. (2011). Applying tests of equivalence for multiple group comparisons: Demonstration of the confidence interval approach. *Practical Assessment, Research & Evaluation, 16*, 1–6.

Rusticus, S. A., & Lovato, C. Y. (2014). Impact of sample size and variability on the power and type I error rates of equivalence tests: A simulation study. *Practical Assessment, Research & Evaluation, 19*, 1–10.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*, 657–680.

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3*, 403.

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning, 33*, 1–19. doi:10.1111/jcal.12172

Sivula, M., & Robson, E. (2015). E-testing in graduate courses: Reflective practice case studies. *MBA Faculty Conference Papers & Journal Articles, 87*, 1–8.

Srikanth, M., & Asmatulu, R. (2014). Modern cheating techniques, their adverse effects on engineering education and preventions. *International Journal of Mechanical Engineering Education, 42*, 129–140. doi:10.7227/IJMEE.0005

Stack, S. (2015). The impact of exam environments on student test scores in online courses. *Journal of Criminal Justice Education, 26*, 1–10. doi:10.1080/10511253.2015.1012173

Tinkelman, D. (2012). Using auditing concepts to discourage college student academic misconduct and encourage engagement. *Journal of Academic and Business Ethics, 5*, 1–28.

Uanhoro, J. (2017). *Effect size calculator*. Retrieved from https://effect-size-calculator.herokuapp.com

Varble, D. (2014). Reducing cheating opportunities in online tests. *Atlantic Marketing Journal, 3*, 131–149.

Walters, A. A., & Hunsicker-Walburn, M. J. (2015). Exploring perceptions of technology's impact on academic misconduct. *Journal of Applied Research in Higher Education, 7*, 32–42.

Wong, S., Yang, L., Riecke, B., Cramer, E., & Neustaedter, C. (2017, September). Assessing the usability of smartwatches for academic cheating during exams. In M. Jones (Ed.), *Proceedings of the 19th international conference on human-computer interaction with mobile devices and services* (p. 31). New York, NY: ACM.

The ***Journal of Social, Behavioral, and Health Sciences*** is an open-access, peer-reviewed, online interdisciplinary journal focusing on research findings that address contemporary national and international issues. Its objectives are to (a) encourage dialogue between scholars and practitioners in the social, behavioral, and health sciences that fosters the integration of research with practice; (b) promote innovative models of interdisciplinary collaboration among the social, behavioral, and health sciences that address complex social problems; and (c) inform the relationship between practice and research in the social, behavioral, and health sciences.

Walden University Publishing: http://www.publishing.waldenu.edu