Walden University
ScholarWorks

Walden Dissertations and Doctoral Studies

Walden Dissertations and Doctoral Studies Collection

2018

Alternative to Proctoring in Introductory Statistics Community College Courses

Lena Feinman
Walden University

Follow this and additional works at: https://scholarworks.waldenu.edu/dissertations

Part of the Instructional Media Design Commons

This Dissertation is brought to you for free and open access by the Walden Dissertations and Doctoral Studies Collection at ScholarWorks. It has been accepted for inclusion in Walden Dissertations and Doctoral Studies by an authorized administrator of ScholarWorks. For more information, please contact ScholarWorks@waldenu.edu.

Walden University

College of Education

This is to certify that the doctoral dissertation by

Yelena Feinman

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee
Dr. Deborah Bauder, Committee Chairperson, Education Faculty
Dr. Wade Smith, Committee Member, Education Faculty
Dr. Gerald Giraud, University Reviewer, Education Faculty

Chief Academic Officer
Eric Riedel, Ph.D.

Walden University
2017

Abstract

Alternative to Proctoring in Introductory Statistics Community College Courses

by

Yelena Feinman

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Educational Technology

Walden University

February 2018

Abstract

The credibility of unsupervised exams, one of the biggest challenges of e-learning, is currently maintained by proctoring. However, little has been done to determine whether expensive and inconvenient proctoring is necessary. The purpose of this quantitative study was to determine whether the use of security mechanisms, based on the taxonomy of cheating reduction techniques rooted in the fraud triangle theory, can be an effective alternative to proctoring. A quasi-experimental 1 group sequential design was used to answer the research questions whether the format, proctored versus unproctored, order in which the exams are administered, course delivery mode, and instructor make a difference in student performance. The archival scores of 850 Californian community college students on 2 sets of equivalent proctored and unproctored web-based exams in face-to-face, hybrid, and online introductory statistics courses taught by 7 instructors were compared. The format effect was tested with repeated-measures ANOVA; the order, course delivery mode and instructor effects were tested with mixed ANOVA. No significant difference in scores in Set 1, and significantly lower scores on unproctored exams in Set 2 indicated that the used security mechanisms allowed for maintaining the credibility of the exams without proctoring. There was no significant difference in scores across the course delivery modes in both sets and instructors in Set 2, but significant order effect was observed. Further research on order effect was recommended. With the use of the utilized security mechanisms education will get an inexpensive and convenient way to increase the credibility of unsupervised web-based exams, and the society will gain more online college graduates with credentials that reflect their knowledge.

Alternative to Proctoring in Introductory Statistics Community College Courses

by

Yelena Feinman

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Educational Technology

Walden University

February 2018

Acknowledgments

My dissertation journey would not have been so productive, interesting, and enjoyable without the chair of my committee, Dr. Deborah Bauder, who has been patiently answering my endless questions with prompt, thoughtful, informative, and wise responses, providing high-quality support and guidance at all levels on an ongoing basis. I would not be able to achieve this goal without the constitutive feedback and knowledgeable recommendations of my methodologist Dr. Wade Smith and the involvement of the university research reviewer Dr. Gerald Giraud. I am very grateful for this. I am deeply thankful for assistance and help of the statistics coordinator, math and science division dean, dean of planning and research, vice president of instructions, and instructional technologist of my college. I am also grateful for being a member of a highly professional team of statistics instructors, without whom this journey would not be so successful. My heartfelt thanks go to my lovely family: my husband, for his unwavering support and understanding of my extreme busyness, and our amazing daughter, for her strict but fair critique of my writing.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to the Study

Web-based instruction has become widespread in higher education. In 2015, 92% of all community colleges in the United States offered at least one online degree (Instructional Technology Council [ITC], 2016). In the same year, about 57 % of all community colleges in the country increased the number of hybrid courses, courses in which up to 70% of instruction take place online (ITC, 2016). More than 70% of the institutions added web-assisted, face-to-face offerings-courses that meet regularly on campus but incorporate the Internet to deliver content, activities, and assessment (ITC, 2016). The growth of web-based instruction brought new advantages and challenges to higher education that required instructors to reconsider many strategies, one of which is student academic assessment (Burke & Bristor, 2017; Ladyshewsky, 2015). New technological advances and the development of learning management systems created favorable opportunities for implementing web-based assessment in all instructional delivery modes (Arnold, 2016; Bain, 2015). However, the same technological advances increased the potential for academic dishonesty, especially during unsupervised exams when students can cheat with ease (Bain, 2015; Burke & Bristor, 2017; Malesky, Baley, and Crow, 2016; Srikanth & Asmatulu, 2014). For this reason, credibility of online assessment is one of the biggest challenges of distance education (Arnold, 2016; Burke & Bristor, 2017; Faurer, 2013; Srikanth & Asmatulu, 2014).

Several procedures, methods, and technologies are available to maintain academic integrity. Among them are cheating-prevention institutional policies, proctoring, the use of different versions of the same test, time restriction, implementation of test questions focused on understanding, and utilization of exam security mechanisms available in most learning management system (LMS) (California Community Colleges Chancellor Office [CCCCO], 2013; Moten, Fitter, Brazier, Leonard, & Brown, 2013; Stack, 2015). The most popular type of exam security is physical proctoring (CCCCO, 2013). However, physical proctoring is time and money consuming for both students and institutions (Cluskey, Craig, & Raiborn, 2012), and may be inconvenient or impossible for some online students. Inability to take an exam online can deny access to higher education of many potential students. Online proctoring, a recent alternative to physical proctoring, is not suitable for many students and institutions due to numerous technological requirements, complicated integration with LMS, high cost, and the time needed for going through remote connections (Atoum, Chen, Liu, Hsu, & Liu, 2015; London, 2014; Milone, Cortese, Balestrieri, & Pittenger, 2017; Pittman, 2015). At some colleges, physical proctoring cannot be required because of the districts' policies, and online proctoring cannot be used due to its cost and technical difficulties (Bain, 2015). Faculty of these institutions, who have concerns regarding cheating, may refuse to teach fully-online courses (Ladyshewky, 2015; Varble, 2014). Thus, out of all available cheating prevention techniques, proctoring is the most frequently used one, but this security mechanism does not satisfy the demands of instructors and their students.

Although academic dishonesty is a problem in higher education regardless of an instructional delivery method (Bain, 2015; Burke & Bristor, 2017; Lang, 2013; McCabe, Butterfield, & Trevino, 2012) and can influence exam scores, cheating was not the topic of my investigation. Instead, this quantitative study was focused on secured, web-based assessment as an alternative to proctoring. If the study shows that secured, web-based exams can be a viable alternative to proctoring, the credibility of online assessment can be sustained during unproctored online exams. Expensive and inconvenient proctoring may be avoided. More students with full-time jobs and family commitments will be able to get degrees; more instructors will be willing to teach fully-online courses. Although an ability to take a web-based exam at a convenient location is especially important for fully-online students, administering secured, unproctored tests in hybrid or face-to-face courses can allow for not spending valuable in-class time on assessment, but rather on learning and instruction.

This chapter provides an overview of the literature regarding the changing landscape of academic assessment and subsequent challenges to web-based testing, describes web-based exam security mechanisms, gives a synopsis of studies that incorporated some security methods and compared scores on proctored and unproctored exams, and identifies gaps in these studies and the need for the given investigation. Then the chapter introduces the purposes of the study, the research questions, the study's theoretical foundation, research design, and methodology. Definitions of key terms, the study's assumptions, scope, delimitations, limitations, and significance conclude the chapter.

Background

Emerging information and communication technologies deeply influence academic assessment (Chen 2014; Desai, 2015; Redecker & Johannessen, 2013). With the rapid utilization of LMS, the traditional pencil-and-paper tests are being gradually superseded by web-based examinations, which are convenient for students and instructors, cost-efficient, and bring new opportunities for teaching and learning (Ladyshewky, 2015; Redecker & Johannessen, 2013; Varble, 2014). Flexibility in creating online tests, building them from scratch, or quickly assembling them using test-banks provided by the publishers has made web-based testing widespread in higher education (Arnold, 2016; Redecker & Johannessen, 2013; Varble, 2014). Paper-free testing, automated scoring of multiple choice and short answers questions, immediate feedback, automatic recording of scores in the gradebook, and immediate basic test items' analysis are other qualities of web-based exams highly valued by educators and administrators (Hameed, 2016; Shute & Rahimi, 2017; Varble, 2014). Computerized adaptive testing, automatic scoring of free response questions, game and simulation-based exams, and interactive test-items are developing attributes of online testing that have the potential to bring all forms of assessment to a new level within the next few years (New Media Consortium, 2015; Kaya & Tan, 2014; Shute & Rahimi, 2017). Thus, technology allowed for many favorable opportunities in academic assessment.

Challenges of Web-Based Assessment

With the development of wireless Internet, social networks, and portable electronic devices, instantaneous access to needed information and its dissemination can occur any time at any location (Kainz, Cymbalak, & Jakab, 2015; Ladyshewky, 2015). The portable electronic devices may include all types of cell phones, earphones, iPads, texting devices, smart-pens, and multifunctional watches and glasses (Bain, 2015; Moore, Head, & Griffin, 2017; Srikanth & Asmatulu, 2014). These advances create favorable conditions for academic cheating during all forms of assessment, but especially in online web-based testing when students have direct unsupervised access to computers and social networks (Moore et al., 2017; Moten et al., 2013; Srikanth & Asmatulu, 2014; Kaya & Tan, 2014).

In 2013, hundreds of state exams were reviewed in California after students posted the pictures of exam questions on Facebook (Davis, 2014). One out of every four college students admitted to using a smart phone for cheating during tests or other assessments (Srikanth & Asmatulu, 2014). Unauthorized use of the Internet during online exams was observed by 40% of instructors at Texas Tech (McCabe et al., 2012; Butterfield & Trevino, 2012). Engineering students were caught cheating by using the online solution manual while taking a web-based exam (McCaslin & Brown, 2015). Indiana University dentistry students were suspended for hacking into computers to access web-based exam answers (Glazer, 2013). Over 1,230 Massive Open Online Courses [MOOC] students copied answers during unproctored certificate exams using multiple online accounts (Northcutt, Ho, & Chuang, 2016).

For the reasons described above, many instructors are not sure whether they can achieve integrity on unsupervised web-based exams comparable with the integrity of pencil-and-paper tests proctored by a human (Chen, 2014; Fask, Englander, & Wang, 2015; Ladyshewsky, 2015; Varble, 2014). Trustworthiness of online exams scores has been a continuing concern of administrators and policy makers since e-learning became widespread (Arnold, 2016; Chen, 2014; Faurer, 2013; Higher Education Opportunity Act [HEOA] Public Law No.110-315, 2008). Thus, institutions with distance education offerings and their online instructors are challenged to find ways in increasing credibility of online exams (Bain, 2015; Chen 2014; Faurer, 2013; Moore et al., 2017; Malesky et al., 2016). These challenges created a need for finding mechanisms that allow for increasing credibility in online environments (Bain, 2015; Malesky et al., 2016; Stack, 2015).

Web-Based Exam Security Mechanisms

In 2009, the Western Cooperative for Educational Telecommunications (WCET), now known as Western Interstate Commission for Higher Education (WICHE) Cooperative for Educational Technologies, in collaboration with the ITC produced a document that lists the best strategies to promote academic integrity (CCCCO, 2013). In addition to California where this study was conducted, the WICHE has 15 more states: Alaska, Arizona, Colorado, Commonwealth of the Northern Mariana, Island Hawaii, Idaho, Montana, Nevada, New Mexico, North Dakota, Oregon, South Dakota, Utah, Washington, and Wyoming (CCCCO, 2013). Many recommendations of this document related to assessment can be used with any instructional delivery mode (McGee, 2013). Adequate institutional policies on academic misconduct, proctoring, and exam questions that require higher order thinking skills are several of the recommendations (WCET, 2009). The document also suggests the use of security mechanisms available in existing learning managements systems.

Colleges and universities with clearly stated institutional academic integrity policies enforced by administration and faculty have a lower rate of academic misconduct in face-to-face and online courses (Bain, 2015; Burke & Bristor, 2017; McCabe et al., 2012). Defining academic dishonesty and indicating consequences of misconduct on the syllabus reduce academic cheating (Bain, 2015; McGree, 2013). In addition to the academic integrity syllabus statement, an announcement about an instructor's ability to view the student logs' activities during online exams may preclude academic dishonesty (Beck, 2014).

Physical proctoring is widely used in higher education (CCCCO, 2013; Pittman, 2015). However, this type of cheating prevention is expensive, inconvenient, or sometimes impossible for many fully-online students and institutions, and goes against the premise of distance education to complete the entire course online (Bandyopadhyay & Barnes 2014; Faire, 2013; Ladyshewsky, 2015). It is not surprising that many online instructors do not use any type of proctoring for their online exams (Bandyopadhyay & Barnes, 2014).

The cognitive level of exam questions may influence student cheating behavior (Ladyshewsky, 2015; Varble, 2014). Test items that measure lower order thinking are easy to look up on the Internet and make cheating more appealing (McGee, 2013; Varble, 2014). Thus, using higher order thinking questions may increase credibility of exams (Ladyshewsky, 2015; Varble, 2014).

Numerous inbuilt LMS tools may reduce the likelihood of cheating on web-based exams (Moodle, 2015; Stack, 2015). Synchronous exams that are open right before the scheduled time minimize dissemination of test items and possible hacking into the system (Moodle, 2015; Moten et al., 2013; Srikanth & Asmatulu, 2014; Stack, 2015). Setting a fixed period precisely needed to answer all exam questions limits or eliminates completely having additional time to text, call a helper, copy the question, and search the Internet or other sources (Ladyshewsky, 2015; Varble, 2014; Stack, 2015). Randomization of exam questions and answer choices, one question per page, and blocked backtracking and feedback options reduce unauthorized collaboration during exams (Beck 2014, Stack, 2015; Varble, 2014). Although the literature suggests some other cheating prevention strategies that can be used during unproctored web-based exams, this study was focused on the security mechanisms described above.

Gaps in the Literature

Numerous previous studies, a complete critique of which is given in Chapter 2, have compared student performance during proctored and unproctored exams utilizing none or some security mechanisms and got different results (Ardid, Gomez-Tejedor, Meseguer-Duenas, Riera, & Vidaurre, 2015; Arnold, 2016; Beck, 2014; Fask et al., 2015; Ladyshewsky, 2015; Sivula & Robson, 2015; Stack, 2015; Varble, 2014)**.** Sivula and Robson (2015) found that graduate students performed 34% better on an online unproctored exam without any security mechanisms, unlimited time, and resources. Similarly, Fask et al. (2015) did not use any security mechanisms and found that students' performance on unproctored exams in an introductory statistics university course was significantly better than on proctored exams (F=7.88, p=.0000028). The latent variable approach used in Fask's et al. (2015) study showed that the latent variable cheating had a direct effect on unproctored exam scores, but not on proctored exam scores, which could indicate the occurrence of cheating on the unproctored exam. However, cheating per se was not the focus of the given investigation.

Beck (2014) incorporated a cheating warning statement, randomization, one question per page, and blocked backtracking and found that there was no significant difference in student's performance on multiple choice unproctored and proctored economics exams (t =.347, p > .05). In addition to all security mechanisms used by Beck (2014), Stack (2015) incorporated synchronous testing and lockdown browser in criminology courses and did not find a significant difference in students' performance during proctored and unproctored exams (b =1.08, p > .05). Ardid et al. (2015) used multiple versions of the test as only one security mechanism and found that student performance on unproctored physics exams was significantly better than on proctored ones (p < .005). While Varble (2014) found that marketing university students did significantly better on unproctored web-based exams than on proctored pencil-and-paper test (F=4.47, p < .01), Ladyshewsky (2015) determined that the postgraduate business students performed better on the proctored exams (p was not stated). The researchers in both studies used the same security mechanisms. However, unlike Varble (2014), Ladyshewsky's exams included mostly high order thinking questions (Ladyshewsky, 2015).

Similar to Beck (2014), Arnold (2016) compared students' scores on proctored and unproctored tests using human capital variables, but with much bigger sample size of 500 students, and found that first year undergraduate economics students performed better on unproctored formative exams (F =1.16; p < .05). The researcher used only two cheating prevention mechanisms: randomization of multiple-choice questions and time restriction (Arnold, 2016). The analysis of these studies, more detailed description of which is given in Chapter 2, identifies several gaps.

The studies described above did not control for exam delivery mode; the researchers administered proctored exams in pencil-and-paper format and unproctored exams in a web-based format. The authors of the articles either payed little attention to equivalency of the proctored and unproctored exams (Ardid et al., 2015; Arnold, 2016) or claimed that the congruence could not be established (Fask et al., 2015). None of the researchers examined the effect of the order in which the exams were administered or instructor's effect. Many of the studies incorporated a few to no cheating prevention mechanisms. Little was done to investigate whether proctored exams would prove unnecessary if appropriate cheating prevention methods were used.

Need for the Study

While academic dishonesty during unproctored exams is an issue (Fask et al., 2015; Shute & Rahimi, 2017; Sivula & Robson, 2015), the given study was not about cheating prevention per se. Previous researchers, utilizing not systematically chosen security mechanisms, compared student performance on proctored and proctored exams and obtained mixed results (Beck, 2014; Ladyshewsky, 2015; Varble, 2014). These mixed results did not allow for determining if inconvenient and expensive proctoring is necessary. There was a need to find ways to increase credibility of unsupervised web-based testing to a level comparable to credibility of proctored exams providing students convenience they need (Ladyshewsky 2015; Stack, 2015). The given study was designed to fulfill this need by investigating whether student performance on proctored and unproctored exams is significantly different when systematically selected security mechanisms are used.

Problem Statement

Credibility of academic assessment is one of the biggest challenges of constantly growing distance education (Arnold, 2016; Faurer, 2013; Malesky et al., 2016; Nash, 2015; Shute & Rahimi, 2017). Specifically, educators and administrators in higher education are concerned that students may use unauthorized help during unsupervised online exams (Fask et al., 2015; Faurer, 2013; Ladyshewsky, 2015; Shute & Rahimi, 2017). In the 2015 survey conducted by the ITC among community colleges, adequate and accurate assessment of students' knowledge and performance in an online environment was listed as the second top challenge after student readiness for e-learning (ITC, 2016). In the same survey conducted in 2016, student assessment and performance in an online environment was listed as the first top challenge, which was followed by student readiness and retention (ITC, 2017).

Physical proctoring is a frequently used security mechanism (Stack, 2015; Lee-Post & Hapke, 2017). However, physical proctoring consumes time and money for both students and institutions and might be inconvenient or impossible for some individuals (Desai, 2015; Ladyshewsky, 2015; Lee-Post & Hapke, 2017). The use of this cheating prevention mechanism in an online environment especially impacts students with extremely busy schedules and students who live far away from proctoring locations (Ladyshewsky, 2015). Remote proctoring is not suitable for many students and institutions due to numerous technological requirements, complicated integration with

LMS, high cost, and the time needed for going through remote connections (London, 2014; Milone et al., 2017; Pittman, 2015). The disconnect between high demand in online learning and inability to maintain credibility of unsupervised web-based exams without inconvenient and expensive proctoring constitutes a problem.

Previous research studies have examined student performance during proctored and unproctored exams (Ardid et al., 2015; Arnold, 2016; Beck, 2014**;** Fask et al., 2015; Ladyshewsky, 2015; Sivula & Robson, 2015; Stack, 2015; Varble, 2014). Some of these studies showed that without any security mechanisms, students’ scores on unsupervised exams were much higher than on proctored exams (Fask et. al., 2015; Sivula & Robson, 2015). The studies in which some security mechanisms were used either found that students' performance on unproctored exams was better (Arid et al., 2015; Arnold, 2016; Varble, 2014) or there was no difference in students' scores (Beck 2014; Stack, 2015**).** The exams in these studies used a few or no apparent systematic secured mechanisms. Moreover, none of these studies utilized web-based format during supervised exams or examined the order in which proctored and unproctored exams were administered. The given investigation addressed these gaps.

Purpose of the Study

The purpose of this quantitative study was to investigate whether inconvenient and expensive proctoring is necessary when web-based exams with systematically selected nonbiometric security mechanisms are used. The relationship between the format in which equivalent automatically-scored, secured, web-based exams were administered, proctored versus unproctored, and exam scores was examined by comparing archived test scores of one group of students during proctored and unproctored exams. The absence of a significant difference between individual student scores would suggest that expensive and inconvenient proctoring is unnecessary. The investigation also examined the effects of order in which proctored and unproctored exams are administered, course delivery modes (a) web-assisted face-to-face, (b) hybrid, and (c) fully-online, and instructors because these extraneous factors may influence the exams' scores (Beck, 2014; Fask et al., 2015; Stack, 2015). The study incorporated a well-developed design and utilized security mechanisms based on the best practices suggested by the literature. Although proctoring itself can be perceived as a security tool, in this study the security mechanisms were synchronous testing, restricted time, blocked backtracking, deferred feedback, randomization, higher thinking levels of exam questions, and policies on academic misconduct. The main independent variable was the exam format: proctored versus unproctored (IV1). The dependent variable was the web-exam score (DV). The order in which proctored and unproctored exams were administered (IV2), course delivery mode (IV3), and the instructor (IV4) were additional independent variables.

Research Questions and Hypotheses

The study was designed to analyze the proctored and unproctored exam scores of community college students in web-assisted face-to-face, hybrid, and fully-online sections of Introductory Statistics taught by a team of instructors who utilized the same curriculum, instructional materials, and assessment. The web-assisted, face-to-face sections are sections that regularly meet on campus but incorporate activities and assessment delivered through Learning Management System (LMS). The hybrid sections have a few mandatory on-campus meetings that involve instructions and up to 79% of content is delivered online (ITC, 2016). The online sections do not have mandatory meetings on campus that involve instruction; 100% of content is delivered through LMS.

In this study, one group of students took two pairs (sets) of proctored and unproctored web-based exams in a certain sequence. In the first set, the proctored exam was followed by the unproctored exam; in the second set, the unproctored exam was followed by the proctored one. The exams had multiple-choice, matching, and short answer questions only, all of which were automatically scored by the LMS Moodle. The content validity and equivalency of the exams within each set and between the sets were established by the experts in the subject matter. More details about the structure and equivalency of the exams are provided in Nature of the Study, and the complete description of these aspects is given in Chapter 3. The construct validity and reliability of the exams are analyzed in Chapter 4.

The study answered the following research questions:

When equivalent automatically-scored web-based exams with the same security mechanisms are used,

RQ1: Is there a relationship between the exam format (IV1), proctored versus unproctored, and student scores (DV)?

> $H_0$1: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms.

H_A1: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms.

RQ2: Is there a relationship between the order (IV2) in which proctored and unproctored exams are administered and student scores (DV)?

$H_0$2: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the order in which exams are administered.

H_A2: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the order in which exams are administered.

RQ3: Is there a relationship between the course delivery mode (IV3), (a) web-assisted face-to-face, (b) hybrid, (c) fully online, and students’ scores (DV)?

$H_0$3: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the course delivery mode.

H_A3: There is a significant difference in students' performance on automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the course delivery mode.

RQ4: Is there a relationship between the instructor (IV4) and students' scores (DV)?

$H_0$4: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the instructor of the course.

H_A4: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the instructor of the course.

Theoretical Framework for the Study

The taxonomy of cheating reduction mechanisms (Varble, 2014) based on the fraud triangle theory (Cressey, 1950) is the theoretical foundation for the study. The department where the given investigation took place carefully selected security mechanisms to reduce cheating during web-based exams. However, the purpose of the study was not necessarily to reduce cheating, but to investigate whether the most common but inconvenient and expensive cheating prevention method, proctoring, is necessary if other security mechanisms are used. This question was answered by comparing students' performance on proctored and unproctored web-based exams. Although there may be many reasons why students can perform differently on exams (Fask et al., 2014), the study framework deals with academic dishonesty because proctoring is used to minimize an impact of possible cheating on student performance (Varble, 2014). The study's theoretical framework explains which security mechanisms can substitute for proctoring and why. For this reason, the selection of the security mechanisms made by the department was explained by the taxonomy rooted in the fraud triangle theory.

An Overview of the Fraud Triangle Theory

Cressey (1950), the author of the trust violation theory, now known as the fraud triangle theory, identified three major factors needed to commit a fraud: *incentive/need, opportunity*, and *rationalization*. These factors were mapped into education settings and used for the understanding, prediction, and prevention of academic cheating (Becker, Connolly, Lentz, and Morrison, 2006; Lewellyn & Rodriguez, 2015; Nkundabanyanga, Omagor, & Nalukenge, 2014; Tinkelman, 2012; Widianingshi, 2013). In education, asynchronous examinations and unlimited time on tests are some factors that may increase the opportunity to cheat (Nkundabanyanga et al., 2014; Tinkelman, 2012). The need to maintain a high GPA and be eligible for scholarships and prestigious universities may stimulate the incentive to cheat (Nkundabanyanga et al., 2014, Lewellyn & Rodriguez, 2015). Students usually rationalize their dishonest behavior claiming that it is not clear what constitutes academic misconduct and no one gets caught (Nkundabanyanga et al., 2014; Tinkelman, 2012). Several researchers tested the fraud triangle theory in educational environments and found that this framework is suitable for research on academic cheating (Becker et al., 2006; Nkundabanyanga et al., 2014; Widianingshi, 2013). The taxonomy of cheating reduction methods with potential to reduce academic dishonesty on online exams was developed based on the findings of these studies and best practices for maintaining academic integrity (Lewellyn & Rodriguez, 2015; Tinkelman, 2012; Varble, 2014).

An Overview of the Taxonomy of Cheating Reduction Methods

The taxonomy 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 opportunities, needs, and rationalization. The opportunity reduction category may involve time restriction, blocked backtracking, and higher order thinking levels of test items (Tinkelman, 2012; Varble, 2014). The need reduction category emphasizes the true value of education and acquired knowledge, importance of the course content for a future profession, assignments that involve students in active learning (Tinkelman, 2012). The rationalization reduction category may include institutional policies, cheating statements on the syllabus (Varble, 2014), and building an atmosphere of appreciative education where instructors and students respect each other (Lang, 2013).

According to the fraud triangle theory, if any of these factors is reduced, neutralized, or blocked, less cheating should take place (Cressey, 1950; Lewellyn & Rodriguez, 2015; Varble, 2014). Therefore, the use of the mechanisms described in the taxonomy of cheating reduction has the potential to reduce academic dishonesty on online tests to a level comparable to proctored pencil-and-paper assessments (Varble, 2014). A more detailed explanation of this theoretical framework and research on it is given in Chapter 2.

Application of the Framework to the Study

Out of the three factors of the fraud triangle, the opportunity is the most problematic in a web-based environment because technology may vastly increase students’ opportunities to cheat (Bain, 2015; Tinkelman, 2012). Additionally, this factor was found to be the most significant determinant of cheating in the literature (Nkundabanyanga et al., 2014). Thus, although the department selected some needs and rationalization reduction methods, the main emphasis was made on incorporation of the opportunity reduction mechanisms. The detailed description of the reduction mechanisms used by the department is provided in Chapter 2.

The hypothesis that was tested is that there is no significant difference in student performance on proctored and unproctored exams when the same security mechanisms are used. The literature suggests that absence of any security mechanisms on unproctored exams increases students’ exam scores significantly (Fask et al., 2015; Sivula & Robson, 2015), which may indicate an occurrence of cheating (Fask et al., 2015). Some previous studies also showed that students may perform better on unsupervised exams if security mechanisms are utilized, but not chosen systematically (Ardid et al., 2015; Arnold, 2016; Varble, 2014). Therefore, systematic selection of cheating prevention mechanisms is necessary (Stack, 2015). The taxonomy of cheating reduction methods describes security mechanisms that have the potential to reduce factors needed for cheating to occur (Cressey, 1950; Varble, 2014). If the chosen web-based exam security mechanisms correspond to the taxonomy’s mechanisms that have the potential to reduce opportunities, need, and rationalization to cheat, academic dishonesty on unproctored exams may be minimized, and the performance on unproctored and proctored exams may be similar.

Nature of the Study

The goal of this quantitative study was to compare individual community college student scores on two pairs (sets) of proctored and unproctored introductory statistics web-based exams. In the first set, the proctored exam was followed 7-10 days later by the unproctored exam; in the second set, the unproctored exam was followed 7-10 days by the proctored one. All four exams involved in the study were part of the regular educational practices of the math department and were administered in accordance with the curriculum: the first set in the middle of the semester, the second set at the end. The exams within each set were alternative tests created by the department to assess the same topics in the curriculum: They had the same questions with different numerical values and themes. The exams between the two sets covered very similar areas in the curriculum and had the same structure, the same number of questions in the same formats, the same level of cognitive and conceptual difficulties, and incorporated the same security mechanisms. The validity and equivalency of the exams within each set and between the sets were established by the professors of the department, who are experts in the subject matter. More detailed discussions on the structure and equivalency of the exams are given in Chapter 3.

A Rationale for the Selection of the Design

The study utilized a quasi-experimental design. A high quality quasi-experimental design is the best and most valid available approach in natural educational settings where randomization usually is not possible or unethical (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002; Thompson & Panacek, 2006). In the community college where the study took place, students are not randomly assigned to exams that are part of a regular educational practice. For this reason, a true experimental design with random assignments, frequently a preferable approach due to its higher internal validity (Cook & Campbell, 1979; Thompson & Panacek, 2006), was not possible. The internal validity of the study's quasi-experimental approach was increased by utilizing a well-developed design structure, which controlled for extraneous variables as much as possible, and carefully-selected data analysis techniques. With the precautions described above, the results of the study might be generalizable for similar institutions.

In the study's design, a single group of students, who were used as their own control, took two sets of proctored/unproctored and unproctored/proctored exams in a certain sequence. In the first set, the proctored exam was followed by the unproctored one; in the second set, the order of the exams was reversed. This design, which is called quasi-experimental single group sequential design (Thompson & Panacek, 2006), controls for potentially confounding factors (Thompson & Panacek, 2006) such as differences in students' backgrounds or instructors' teaching styles, and requires fewer subjects (Thompson & Panacek, 2006). A synopsis of the design's structure begins with the description of the involved variables.

An Overview of the Key Variables

The main independent variable is the format (IV1) of the web-based exams. Although proctored exams can be perceived as a cheating prevention mechanism, in this study proctored is the value of the independent variable exam format. The second value of this variable is unproctored. The web-based exam score (DV) is the dependent variable. The order (IV2) in which the proctored and unproctored web-based exams are administered, the course delivery mode (IV3), which has three categories: (a) web-assisted face-to-face, (b) hybrid, and (c) fully online, and the instructor (IV) of the course are additional independent variables.

An Overview of the Methodology

The investigation took place at a suburban Californian community college that serves about 9,000 students every semester and offers transfer programs in 11 subject areas. Nine out of these 11 subject areas have Introductory Statistics as a transfer requirement. Art and Foreign Languages and Pure Mathematics, Engineering, and Computer Science constitute the remaining two subject-areas. Thus, around 300 students take Introductory Statistics every term. More detailed description of the study's setting is given in Chapter 3.

All students who were enrolled in web-assisted face-to-face, hybrid, and fully-online sections of Introductory Statistics at the college where the investigation took place constituted the study's population. Because the scores of each student in this population were analyzed, a census sample was utilized. Thus, the study's findings can be generalized to web-based courses offered in similar institutions.

The math department of the college has implemented a procedure for administering web exams in proctored and unproctored formats in all web-based introductory statistics classes. As part of the institutional student learning outcomes analysis and data collection for educational purposes, the department has been collecting and analyzing all students' exam scores and demographics. Some of these archived data were used in the study. The web-based sections of the course are taught by a team of instructors who use the same textbook, web materials, assignments, and administer the same exams during the same week. Although the course curriculum includes several web-based exams, to compare the individual student performance on proctored and unproctored web-based exams and investigate the order, course delivery mode, and instructor effects, the study focused on the analysis of the scores of the following two sets of equivalent exams:

Set 1 (proctored first, unproctored second): In the middle of the semester, the students took the proctored in class secured, web-based exam. In the 7-10 days after that, the same students completed the alternative unproctored online exam. The alternative exam had the same questions and structure, but different numbers. The difference of the individual scores within Set 1 exams was analyzed.

Set 2 (unproctored first, proctored second): At the end of the semester, the same students took the unproctored online exam first and then, 7-10 days later, the alternative proctored web-based exam. The exams in Set 2 also had the same structure and questions, but different numerical values. The difference of the individual scores within Set 2 exams was analyzed.

Since all exams involved in the study were parts of a regular educational practice, 2/3 of the questions in Set 2 covered slightly different areas in the curriculum than the questions in Set 1. However, a team of instructors, experts in the subject matter, established the equivalency of the exams between the two sets. Thus, the exams in Set 1 and Set 2 had the same cognitive and conceptual levels of difficulty, the same structure, the same number of multiple-choice, matching, and short answer questions, and the same security mechanisms. The students did not know that the exams within each set and between two sets were equivalent. To analyze the order effect, the difference of the individual scores of Set 1 exams was compared with the difference of the individual scores of Set 2 exams. An absence of a significant difference in scores between Set1 and Set 2 would indicate that there was no order effect; a significant difference would demonstrate the effect.

All exams incorporated only automatically scored multiple- choice, automatically scored matching, and automatically scored short answer questions, which controlled for grading effect. The differences in numerical values and themes, in addition to randomization of exam questions, reduced the testing effect. The number of questions in all exams and allocated test time were also the same. The use of the same curriculum and the same instructional materials minimized the instruction effect. All data needed for the analyses were provided by the Introductory Statistics coordinator of the math department and presented in a recoded spreadsheet with all identifying information removed to protect the identities of the students and instructors. To increase the power analysis of the study and achieve sufficient size, the data accumulated from Fall 2015 to Summer 2017 semesters were requested. The same four exams involved in the study were administered during these semesters.

The Data Analyses

The number of students who took Set 1 exams differed from the number of students who took Set 2 exams due to attrition. For this reason, to test the hypothesis of the first research question, whether there is no significant difference in students' performance on equivalent unproctored and proctored introductory statistics web-based exams, a one-way repeated-measures ANOVA was applied to Set 1 scores and Set 2 scores separately. A repeated-measures ANOVA is appropriate for the first research question because the variable of interest, the exam score (DV), is measured two or more times (Field, 2013; Hopkins, 2008). A mixed ANOVA was incorporated to test the order, course delivery mode, and instructor effects addressed in the second, third, and fourth research questions. A mixed ANOVA is appropriate for testing these effects because this technique allows for analysis of within-subject and between-subject variables simultaneously (Field, 2013): the independent variables the format (IV1) and order (IV2) are within-subject variables while other two independent variables course delivery mode (IV3) and instructor (IV4) are between-subject variables. More detailed description of the methodology is provided in Chapter 3.

Definitions

The list of the terms and operational definitions related to the study is provided below:

Automated scoring: Automated scoring is a computerized evaluation of responses to multiple-choice, matching, fill-in-the-blank, and short answer questions available in most learning management systems (Scalise & Gifford, 2006).

Alternative exams: Alternative exams are exams that cover the same topics, utilize the same type of questions and security mechanisms, but have different themes and numerical values (Benedict & Zgaljardic, 1998).

Equivalent exams: Equivalent exams are exams that have the same purpose, structure, and level of difficulty (Cummings, 2003). In this study, equivalent exams are exams that have the same security mechanisms, the same time restriction, and include the same type and number of questions on the same or very similar topics, at the same level of conceptual and cognitive difficulties. The equivalent validity of these exams was established by the experts in the field.

Exam format: Exam format is the main independent variable which represents a format of exam administration: proctored versus unproctored (Harmon & Lambrinos, 2008).

Exam score: An exam score is a web-based exam score automatically assigned by the LMS (Harmon & Lambrinos, 2008).

Extraneous variables: Extraneous variables are variables that are not major variables under an investigation, but could have an effect on the dependent variable or on the relation between the independent and dependent variables (Fraenkel, Wallen, & Hyun, 1993).

Fully-online courses: Fully-online courses are courses with no mandatory face-to-face meetings that involve instruction; all content is delivered online (ITC, 2016).

Hybrid courses: Hybrid courses are courses with mandatory on-campus meetings that involve instruction; up to 79% of content is delivered through learning management system or other means of web-based technology (ITC, 2016).

Online proctoring: Online proctoring refers to monitoring of an online exam through a webcam and may include any automated procedures that help to authenticate examinees and ensure security of the exam (Foster & Layman, 2013).

Proctored exams: Proctored exams are exams that are taken under supervision of the instructor or another designated individual approved by the college (Harmon & Lambrinos, 2008).

Unproctored exams: Unproctored exams are unsupervised web-based exams that can be taken at any location with the Internet access (Harmon & Lambrinos, 2008).

Security mechanisms: Security mechanisms are synchronous administering of the exams, restricted time, randomization of exam questions, one question at a time, blocked backtracking, deferred feedback (Stack, 2015), higher cognitive level of exam questions (Ladyshewsky, 2005), cheating warning statements, and institutional policies on academic misconduct (Varble, 2014).

Short answer exam questions: Short answer questions are web-based automatically-scored questions, which require inserting a numerical value, a word, or a short phrase (Daniel & Broida, 2004)

Web-assisted courses: Web-assisted courses are face-to-face courses in which up to 29% of content is delivered through learning management system or other means of web-based technology (ITC,2016).

Web-based courses: Web-based courses are courses in which the entire course content or part of it is delivered through a learning management system (ITC,2016).

Web-Based exams: Web-based exams are exams that are delivered through LMSs (Beck, 2014).

Assumptions

The given study is based on several assumptions. It was assumed that Introductory Statistics students, whose exam scores was analyzed in the study, represented typical community college population. Generalization of the study's results for similar institutions was made based on this assumption. It was also assumed that the use of the same course calendar, the same syllabus, the same textbook and other course materials, the same homework and the same exams eliminated the instruction effect. This assumption increased the internal validity of the study. It was believed that the students took unproctored exams in a quiet environment. Another assumption was related to the quality of archival data. It was assumed that the Introductory Statistics coordinator downloaded and combined the exams scores with demographic information correctly.

It was believed that the instructors proctored all supervised exams well. Because the participants were from the same location and were approximately the same age, it was assumed that possible maturation occurred at the same pace for all involved individuals. Lastly, it was believed that all students who took the unproctored exams were the same individuals who took the proctored exams. The study's focus on nonbiometric security mechanisms was related to the last assumption.

Scope and Delimitations

In the given study, I investigated whether carefully selected security mechanisms can maintain credibility of unproctored web-based exams at the level comparable with credibility of equivalent proctored web-based exams. This specific focus was chosen because unproctored web-based assessment has become widely spread in higher education institutions (Allen, Seaman, Poulin, & Straut, 2016; Arnold, 2016), while commonly used expensive and inconvenient proctoring has not been satisfying the demands of many distance education programs and their students (Ladyshewsky, 2015; Milone et al., 2017). The setting of the future investigation is one of these institutions.

The study was conducted at a suburban community college in California, the math department of which incorporated web-based proctored and unproctored exams in the curriculum of the introductory statistics course offered in web-assisted face-to-face, hybrid, and fully-online modes. While unproctored web-based exams are less expensive than proctored, convenient for students and institutions (Ladyshewsky, 2015), allow for automatic scoring (Stack, 2015), and can save valuable for face-to-face instruction in-class time (Sivula & Robson, 2015), the instructors and administrators of the department were concerned about credibility of unsupervised online tests. The utilization of carefully selected mechanisms based on the best existing practices was suggested to address this issue. In my study, I examined whether the selected security mechanisms can be an effective alternative to proctoring.

Delimitation of the Study

The study was delimited to students who took introductory statistics web-based courses, which are courses offered in web-assisted face-to-face, hybrid, and fully-online modes. Introductory Statistics is a community college course that satisfies transfer and associate degree requirements in 82% of subject areas offered by the college. For this reason, this group of students represented the body of the transferring students of the college well. Additionally, it is the only course in the department that has been offered in web-assisted face-to-face, hybrid, and online formats on a regular basis during the last 5 years. All sections of this course are scheduled in a computer classroom that allows for use of technology needed for proctored web-based exams.

The purpose of the investigation drove the choice of the study's theoretical framework. The goal orientation theory (Alt & Geiger, 2012; Zito & McQuillan, 2010), item response theory (Champlain, 2010; Templin, 2016), and taxonomy of cheating prevention mechanisms (Varble, 2014) based on the fraud triangle theory (Cressey, 1950) are theoretical frameworks related to academic cheating on exams (Alt & Geiger, 2012; Varble, 2014; Wollack, Cohen, & Eckerly, 2015; Zito & McQuillan, 2010). Out of these three theories, the taxonomy of cheating prevention mechanisms is the most appropriate theoretical foundation for the given study because it explains which security mechanisms have the potential to reduce factors needed for cheating to take place (Cressey, 1950; Varble, 2014). If the factors needed for cheating are minimized, student performance on proctored and unproctored exams may be similar and proctoring may not be necessary. The goal orientation theory, which informs why academic dishonesty occurs (Alt &

Geiger, 2012; Cheung, Wu, & Huang, 2016; Zito & McQuillan, 2010), is not suitable because I did not investigate the reasons for cheating. The item response theory, rooted in the computational analysis of students' responses, may be efficient in cheating detection (Shu, Henson, & Luecht, 2013; Wollack, Cohen, & Eckerly, 2015), which was not the focus of the study. For these reasons, the theoretical framework of the study is delimited by the taxonomy of cheating prevention mechanisms based on the fraud triangle theory. The detailed analysis of the theories described above is given in Chapter 2.

The selection of the cheating prevention mechanisms was restricted to nonbiometric security mechanisms. Biometric security mechanisms were not available at the college where the study was conducted. Moreover, although the instructors of the department have reported several occurrences of cheating during unproctored and proctored exams, all these incidences were unrelated to authentication cheating. It was assumed that authentication misconduct was not common among community college students.

Generalizability of the Study

Although the study was conducted with the students' scores on secured unproctored and proctored web-based exams in introductory statistics courses, its results may be generalized to other similar courses. Additionally, because the selected security mechanisms are universal and available in most currently used LMSs, the findings of the investigation may be applicable to any community college with a LMS platform. However, the generalizability might be limited to institutions comparable with community colleges due to possible different student population, class size, and curriculum in other higher education settings.

Limitations

The students were not randomly enrolled in their classes because the enrollment procedure established by the college is based on self-selection. This nonrandom selection could bring threats to external validity and decrease the study's generalizability (Shadish et al., 2002; Slavin, 2008). However, the results of carefully designed studies with nonrandom selection may be generalizable for similar institutions (Shadish et al., 2002; Slavin, 2008; Thompson & Panacek, 2006). The students also were not randomly assigned to proctored and unproctored exams. This limitation could bring threats to internal validity through selection bias due to possible differences in initial characteristics of the participants (Shadish et al., 2002; Slavin, 2008). The study's quasi-experimental single group sequential design, in which each student is used as his or her own control, rules out the selection bias described above (Thompson & Panacek, 2006). Testing effect, fatigue effect, and attrition effect, other potential threats to internal validity (Campbell & Stanley, 1963; Shadish et al., 2002), are discussed in Chapter 2 and Chapter 3.

Unnoticed or uncontrolled extraneous variables can also bring threats to internal validity (Shadish et al., 2002). The study's design controlled for students' initial differences, course content, structure, form, and content of assessment, security mechanisms, and grading. Other extraneous variables, the order in which the exams were administered, course delivery mode, and instructor were analyzed. To reduce threats to construct validity of the study, clear operational definitions are provided. A detailed analysis of the validity threats and how they were addressed is given in Chapter 3.

Limitations of the Study Related to Methodological Weaknesses

I utilized archival data that could bring some limitations due to possible incompleteness of the data set and inability to address certain important aspects of the study (Jones, 2010). The incompleteness of the data set can result in a not large- enough census sample size and small statistical power. To achieve a size required for sufficient statistical power, the students' scores from several semesters were used. The detail description of the obtained data set is provided in Chapter 4.

Biases that Could Influence Study Outcomes

The use of quasi-experimental approach in the study could result in design bias. However, well-developed quasi experiments can have internal validity and accurate results comparable with randomized experiments (Shadish et al., 2002). Thus, the potential study's design biases were lessened by reducing threads to validities described above, controlling for extraneous variables, and implementing adequate statistical methods. The experimenter and response biases were minimized by the use of archived data and involvement of seven instructors; possible grading bias was reduced by automatic scoring. Further analysis of the study's biases is provided in Chapter 3.

Significance

Online web-based assessment has become wide spread in higher education (Allen & Seaman, 2015). However, educators, administrators, and policy makers are concerned that students can cheat with ease during unsupervised exams (Arnold, 2016; Burke &

Bristor, 2017; Fask et al., 2015; Ladyshewsky, 2015). Currently, expensive and inconvenient proctoring is used as the main mechanism to maintain the credibility of online exams (Bandyopadhyay, Barnes, & Bandyopadhyay, 2015; CCCCO, 2013). But little has been done to investigate whether the integrity of online testing can be maintained without proctoring (Stack, 2015). The given study is significant because it was designed to fill this gap by examining student performance on automatically-scored web-based proctored and automatically-scored web-based unproctored exams with carefully selected security mechanisms. To the best of my knowledge, this was the first study on this topic conducted with community college students. Moreover, it was the first study in which the order effect was examined in a natural educational setting.

Contribution of the Study that Advances Web-based Cheating Prevention Practices

A carefully selected combination of existing nonbiometric security mechanisms was studied in the given investigation. If no significant difference in student scores during unproctored and proctored web-based exams with this combination of the security mechanisms is present, online education will gain a powerful cost-efficient alternative to proctoring that can advance existing cheating prevention practices. Although the study was focused on utilization of security mechanism during web-based unproctored exams in introductory statistics community college courses, the results of the study may be applicable to any subject offered in any format.

Implication for Social Change

If the study results suggest that the combination of the security mechanisms may effectively substitute for proctoring, online programs will be able to maintain the integrity of their courses while providing the students the convenience they need. Technology in the form of web-based exams can allow for administering secured unproctored exams in face-to-face or hybrid courses outside of class, reserving in-class time for working on additional concepts, problem- solving, or projects (Sivula & Robson, 2015). The credibility and convenience of online education will increase, more students will attend distance education institutions, the entire society will gain more college graduates with a high potential of becoming valuable professionals in their fields.

Summary

The study about secured proctored and unproctored web-based exams in community college introductory statistics courses is introduced in this chapter. New technological advances brought new challenges to distance education, one of the biggest of which is the credibility of unsupervised exams (Fask et al., 2015; ITC, 2017; Ladyshewsky, 2015). To overcome this challenge, many higher education institutions use inconvenient and expensive proctoring (Bandyopadhyay et al., 2015; CCCO, 2013; Desai, 2015). Numerous previous studies have compared student performance on proctored and unproctored exams **(**Arnold, 2016; Beck, 2014**;** Fask et al., 2015; Stack, 2015), but none of them have examined the combination of systematically chosen security mechanisms used by the department where the study took place.

This study’s purpose was to investigate whether inconvenient and expensive proctoring is necessary to maintain credibility of unsupervised web-based exams. The archived students’ scores on equivalent automatically scored proctored and unproctored web-based exams with the same security mechanisms were compared. The students’ performance on these exams was analyzed with respect to the order in which the proctored and unproctored exams were administered, course delivery mode, and instructor. The potential effectiveness of carefully security mechanisms used during the web-based exams is explained by the fraud triangle theory and taxonomy of cheating prevention mechanisms. The study utilized quasi-experimental single group sequential design in which each student is used as his or her own control. A one-way repeated-measures ANOVA and mixed ANOVA were used to answer the study's research questions. If the results show that there is no significant difference in scores on proctored and unproctored secured, web-based exams, credibility of online assessment may be maintained without inconvenient and expensive proctoring.

An overview of current research on web-based academic assessment, challenges associated with online testing, best practices of cheating prevention techniques, security mechanisms, and the taxonomy of cheating prevention methods based on the fraud triangle theory are provided in Chapter 2. A review of the literature on the impact of proctored and unproctored exams on student performance, and the potential influence of the exams' order, course delivery modes and the instructor on student scores is also given in this chapter. Chapter 2 is concluded with the summary of major themes and gaps found in the literature and the discussion of how the given study filled these gaps.

Chapter 2: Literature Review

Web-based assessment has become widespread in higher education due to the rapid development of learning management systems and other technological advances (Allen et al., 2016; Arnold, 2016; Bain, 2015; ITC, 2017). However, the same advances increased the potential for academic misconduct in all forms of assessment, but especially during unsupervised tests (Corrigan-Gibbs, Gupta, Northcutt, Cutrell & Thies, 2015; Malesky et al., 2016; Nilsson, 2016; Shute & Rahimi, 2017; Srikanth & Asmatulu, 2014). Thus, credibility of unproctored exams became one of the biggest challenges of online programs (Arnold, 2016; Bain, 2015; Faurer, 2013; Ladyshewsky, 2015; Malesky et al.; 2016; Shute & Rahimi, 2017).

Many methods, procedures, and technologies are available to maintain academic integrity (Bain, 2015; Burke & Bristor, 2017; Moten et al., 2013; Srikanth & Asmatulu, 2014). Cheating-prevention institutional policies, proctoring, the use of test questions focused on understanding and interpretation, and utilization of exam security mechanisms available in many LMSs are some of them (Arnold, 2016; Bain, 2015; CCCCO, 2013; Shute & Rahimi, 2017; Stack, 2015). Physical proctoring is a frequently used method of exam security (CCCCO, 2013; Bain, 2015). However, physical proctoring is expensive, time consuming, and inconvenient or impossible for both students and institutions (Ladyshewsky, 2015). Many potential students may be unable to get a degree due to their inability to take physically proctored exams. Remote proctoring, which is still in a stage of development, is costly, not user-friendly, and may raise a violation of privacy issue (Barnes & Paris, 2013; Hylton, Levy, & Dringus, 2016; Milone et al., 2017).

The purpose of the given study was to determine whether inconvenient and expensive proctoring is necessary when carefully selected nonbiometric security mechanisms are used. The investigation was designed to examine the relationship between the format in which equivalent automatically-scored secured, web-based exams are administered, proctored versus unproctored, and exam scores. If there is no significant difference in students' performance on proctored and unproctored exams, inconvenient and costly proctoring may be avoided. Student performance on proctored and unproctored exams can be influenced by the order in which the exams are administered (Fask et al., 2015), by the course delivery mode (Beck, 2014), and the instructor of the course (Beck, 2014; Ladyshewsky, 2015; Stack, 2015). For this reason, the order, course delivery mode, and instructor effects on exam scores was also analyzed.

The challenges of assessment in the digital era are reflected in the current literature. McCaslin and Brown (2015) reported the use of online solution manual by engineering students during a web-based exam. Sivula and Robson (2015) found that graduate students' performance on an unproctored exam without utilization of any security mechanisms was 34% better than during proctored one. Corrigan-Gibbs et al. (2015), who studied cheating in massive open online courses (MOOCs), determined that about 25% of 409 students used the Internet during online exams. Northcutt et al. (2016) found that over 1,230 MOOC students used multiple online accounts to copy answers during unproctored certificate exams.

For the reasons described above, online instructors are challenged to find ways in increasing credibility of online exams (Arnold, 2016; Stack 2015; Varble, 2014).

Weatherly, Jennings, and Wilson (2015) found that about 83% of 88 online instructors who participated in the study were concerned about academic integrity in their classes. About 40 % of the respondents required their students to take proctored exams with a human proctor (Weatherly et al., 2015). However, physical proctoring is expensive and inconvenient for many online students.

In Barnes and Paris's (2013) survey-research about what online instructors are doing to maintain academic integrity, 83 % of the respondents thought physical proctoring was not suitable for many of their distant students due to their geographic location. Almost 90% of the same participants responded that remote proctoring would not be a good choice either due to its cost, user dissatisfaction, privacy issue, and possible lack of effectiveness (Barnes & Paris, 2013). Anderson and Gades (2017) studied online students and their instructor's perceptions of proctoring and found that 93.7% of surveyed students would not take an online course if the cost of physical or remote proctoring is not covered by the institution or online program. The students noted inconvenience of physical proctoring and additional technical requirements of remote proctoring. The instructor listed several difficulties associated with remote proctoring: start up learning curve for instructors and students, students' concern about privacy, the need to provide additional set of instructions to students, additional risk that something can go wrong, and a long waiting period for receiving cleared exams from the remote proctoring companies (Anderson & Grades, 2017). Thus, both students and instructors find proctoring expensive and inconvenient.

This chapter begins with the description of the literature search strategies, which is followed by the description of the study’s theoretical framework, the taxonomy of cheating prevention methods based on the fraud triangle theory, its origin, main propositions, previous and current use in research. The current literature on web-based academic assessment, challenges associated with it, best practices of cheating prevention techniques, and study’s methodology are discussed in the next section. The analysis of the literature about the influence of exam format, proctored versus unproctored, order in which exams are administered, course delivery mode and instructor on student performance is also included in this section. The description of the literature related to the study’s research questions, major themes, and the discussion on how the given study fills the gaps found in the literature conclude the chapter.

Literature Search Strategies

A search of electronic databases available through Walden library and Google Scholar has been conducted to find research articles about web-based assessment, challenges associated with it, existing best practices to overcome these challenges, and related theories. Education Research Complete, ERIC, ResearchGate, SAGE, ScienceDirect, Taylor and Francis Online, and Thoreau are specific databases have been used in the search. The general terms *online assessment*, *cheating on exams*, *web-based testing* were used at the beginning of the process. To narrow the focus, the phrases *cheating prevention methods*, *type of web-based questions*, *Bloom’s Taxonomy in online testing*, and *theories of web-based testing* were incorporated. The main focus of the literature search has been on peer-reviewed full-text articles published from 2012 to

2017. When the study's theory, the fraud triangle theory, was identified, the search for the articles, which utilized this theory in education, was extended before 2012.

The strategies described above have resulted in about 300 articles closely-related to the topic of the investigation, its design, and theoretical frameworks. In addition to peer-review articles, the books on academic cheating by McCabe (2012) and Lang (2013) have been read. Seven dissertations found in ProQuest database have been used as supplemental sources.

Theoretical Foundation

The hypothesis that was tested is that there is no significant difference in student performance on proctored and unproctored equivalent exams when the same security mechanisms are used. The purpose of the given investigation is not necessarily to reduce cheating, but to examine whether inconvenient and expensive cheating prevention technique proctoring is necessary if other security mechanisms are incorporated. However, the theoretical framework of the investigation, the taxonomy of cheating reduction techniques (Tinkelman, 2012; Varble, 2014) rooted in the fraud triangle theory (Cressey, 1950), is directly related to cheating prevention because it informs which security mechanisms can substitute for proctoring and why.

The Origin of the Fraud Triangle Theory

The trust violation theory, which is now known as the fraud triangle theory, was developed by Cressey (1950) who conducted a study, the main purpose of which was to determine why employees violate trust and commit fraud. The researcher interviewed 250 criminals who violated trust and found that a person commits fraud when a combination of the following three factors is present: financial or emotional problem, opportunity to commit a violation, and rationalization by a violator (Cressey, 1950). Opportunity and need are external factors, while rationalization comes from the individual. Based on his study's results, Cressey (1950) concluded that the problem generates the need to commit fraud; favorable conditions for committing a trust violation and unlikeliness to be caught provide perceived opportunity; rationalization enables fraudsters to perceive the crime as acceptable because they see themselves as honest individuals who are in a desperate situation. Cressey's (1950) fraud triangle theory was mapped into education and became an adequate framework for explaining fraud in academia (Becker et al., 2006; Hayes, Hurtt, & Bee, 2006; Lewellyn & Rodriguez, 2015; Malgwi & Rakovski, 2008; Nkundabanyanga et al., 2014).

Mapping the Fraud Triangle Theory into Education

Although Cressey (1950) developed the theory studying financial fraud, all three components of the triangle appeared to be well-suited for understanding academic misconduct (Becker et al. 2006; Malgwi & Rakovski, 2008). Students are expected to pursue their degrees with academic integrity (Malgwi & Rakovski, 2008). However, different forms of academic pressure, for example, the need to maintain a high GPA or be eligible for scholarships, may generate a need to commit academic fraud (Becker et al. 2006; Malgwi & Rakovski, 2008). When the pressure becomes very strong, students look for opportunities to cheat. If the opportunities are present, for instance, instructors use the same exams every semester or do not pay attention to what students are doing during tests, the students go through rationalization, a process of justifying the fraudulent behavior. Students may rationalize their dishonest behavior by claiming that the instructor did not explain what constitutes academic misconduct and none of my friends got caught. The need becomes connected with perceived opportunities through rationalization and academic dishonesty occurs (Becker et al. 2006; Malgwi & Rakovski, 2008).

Malgwi and Rakovski (2008) extended Cressey's (1950) fraud triangle theory by adding specific to education academic fraud risk factors. The risk factors, which were selected based on the previous literature (Bolin, 2004; Hayes et al., 2006), interviews, and pilot surveys conducted by the researchers, were divided into three categories that corresponded to the three elements of the fraud triangle (Malgwi & Rakovski, 2008). The need category included peer pressure and fear of failing the course, losing financial aid or parents' support, and inability to attend prestigious universities and obtain high-paying jobs due to low grades (Malgwi & Rakovski, 2008). The use of electronic devices during exams, absence of supervision and other actions that deter cheating, and easy access to prohibited materials belonged to the opportunities category (Malgwi & Rakovski, 2008). Students' perceptions that everyone who studies hard or pays tuition deserves to pass, that the course is difficult and the instructor is a hard grader, that there are no clear policies on academic misconduct, and that fraud exists everywhere in our society were classified as rationalization risk factors (Malgwi & Rakovski, 2008). Thus, academic cheating risk factors were linked to the three factors of the fraud triangle theory.

Lewellyn and Rodriguez (2015) expanded the fraud triangle further by adding the components that are focused on the fraud act itself. The researchers suggested that the academic fraud act, its execution and its methodology, should be taken into consideration to identify the most effective antifraud techniques. Lewellyn & Rodriguez recommended future research to focus on this aspect of the fraud triangle. Thus, all components of the fraud triangle theory are important in selection of effective cheating prevention techniques.

Adequacy of the Fraud Triangle Theory for Explaining Academic Fraud

Several empirical studies showed that the fraud triangle theory is an adequate framework for understanding, prediction, and prevention of academic dishonesty (Becker et al., 2006; Malgwi & Rakovski, 2008; Nkundabanyanga et al., 2014; Walters & Hunsicker-Walburn, 2015). Becker et al. (2006) tested the fraud triangle theory in educational setting with 476 Midwestern university business students. The researchers used an inability to get the desirable grades and be acceptable by prestigious universities as need measures, when instructors do not deter cheating as opportunity measures, and when the instructor's grading policies are unfair and it is not explained what constitutes cheating as rationalization measures. The results of their investigation showed that the elements of the fraud triangle, need, opportunity, and rationalization, were significant determinants of student cheating (R^2 =.2042, $p \leq .004$). Becker et al. referred to cheating as any form of academic misconduct. The regression model used for the analysis indicated that cheating behavior increased when the need, opportunity, and rationalization increased. Becker et al. concluded that understanding of the fraud triangle elements may help educators to reduce cheating by explaining value of knowledge, improving test security, and clearly stating what constitutes academic misconduct. The authors noticed that although all three factors of the triangle should be considered, educators may have little control of the need and rationalization factors, but can significantly reduce opportunities. The researchers also suggested conducting future research on academic dishonesty grounded in the fraud triangle theory (Becker et al., 2006).

Choo and Tan (2008) tested the robustness of Becker et al.'s (2006) results in their study with 182 accounting and business students in a large public university in California. Like Becker et al., Choo and Tan found that each element of the fraud triangle significantly influenced students' propensity to cheat (p < .01), where cheating was any dishonest behavior that helps students to get an undeserved grade. In addition to Becker et al.'s findings, a full-factorial within-subject ANCOVA in Choo and Tan's study showed a significant three-way *need* x *opportunities* x *rationalization* interaction effect (p = .016), which demonstrated that all three factors affect the students' propensity to cheat. Choo and Tan concluded that their findings indicated that faculty should take into consideration all three factors trying to neutralize the need and rationalization and reduce opportunity to cheat. The authors suggested that to understand which opportunities prevention and need and rationalization neutralization techniques are effective, future research should study the relationship between the triangle's elements in more details.

Malgwi and Rakovski's (2008) investigation was focused on the relationship between the elements. They surveyed 740 students at a large business university in the Northeast to examine which of the fraud triangle elements need, opportunit*y*, and rationalization is the most important determinant of possible occurrence of cheating (Malgwi & Rakovski, 2008). The researcher used their fraud triangle risk factors described above to measure the importance of the elements of the triangle in determining whether to cheat or not (Malgwi & Rakovski, 2008). Cheating in Malgwi and Rakovski's study was any behavior that results in gaining an unfair advantage in academic performance. The researchers utilized a nonparametric Pearson Chi-square test and found that there was a significant difference across the three categories ($\chi^2 = 435.3$, $p = .0001$). The factor analysis showed that about 70% of students rated the need category as the most influential element of the fraud triangle: failing the course and losing financial aid were the top two risk factors. The second influential element was opportunity selected by 20% of all responders: acquiring needed information from a friend or by using electronic devices were the two top risk factors in this category. Only 10% of the participants selected rationalization. Thus, rationalization was the least influential element of the triangle with the top two risk factors: students who study hard deserve to pass and the concepts covered in the course are too hard to learn. These results matched Cressey's (1950) findings according to which the need was the most influential factor followed by the opportunities and then rationalization. Like Becker et al. (2006), Malgwi and Rakovski noticed that, out of the three elements of the fraud triangle, faculty can control opportunities the most.

Further research on the fraud triangle model in education and the relationship between the triangles' elements was conducted by Nkundabanyanga et al. (2014). The researchers examined the effect of the fraud triangle elements with 471 undergraduate and postgraduate university business students in Uganda, Africa (Nkundabanyanga et al., 2014). Unlike Becker et al. (2006) , Choo and Tan (2008), and Malgwi and Rakovski

(2008) who investigated the effect of the triangle’s factors on any type of academic misconduct, Nkundabanyanga et al. focused specifically on cheating during exams. The researchers examined the relationship between each element need, opportunity, and rationalization and academic dishonesty on tests. By using cross-sectional design and structural equation modeling, Nkundabanyanga et al. found that the fraud triangle is a relevant theoretical framework for understanding cheating during exams: the factors of the triangle significantly predicted academic misconduct (R^2 =.36, p = .001). However, opportunit*y* was the most important element which accounted for significant variations in academic cheating on exams (p =.001). The researchers also found that opportunities accounted for significant variations in rationalization (p =.001) and recommended to study the relationship between these two elements further (Nkundabanyanga et al., 2014).

Walters and Hunsicker-Walburn (2015) focused on the relationship between the opportunities and rationalization in their qualitative study on faculty and students’ perceptions of technology’s impact on academic dishonesty. The researchers, who interviewed 40 state university instructors experienced in teaching with technology in face-to-face and online courses and 20 of their students, found that faculty, noticing that technology increased occurrence of cheating, tried to reduce opportunities to cheat by restricting the use of technology. However, these actions resulted in increase of student rationalization: Students felt mistrust, their cheating behavior was justified, and they looked for new ways to cheat undetected. The researchers concluded that reduced opportunities can increase rationalization (Walters & Hunsicker-Walburn, 2015).

All studies described above showed that the fraud triangle theory is a suitable framework for explaining academic dishonesty. Although students may perceive the need as the most influential element (Malgwi & Rakovski's, 2008), faculty can control the opportunity to cheat the most (Becker et al., 2006; Malgwi & Rakovski, 2008; Nkundabanyanga et al., 2014). For this reason, the best cheating prevention practices are mostly focused on exam security mechanisms that reduce opportunity to cheat (Varble, 2014). However, it is also important to consider other elements of the triangle because all three factors independently (Nkundabanyanga et al., 2014) and combined (Choo &Tan, 2008) influence students' propensity to cheat. That is why, in addition to opportunity reduction techniques, methods that neutralize the need and rationalization should be incorporated as well (Tinkelman, 2012; Varble, 2014). Moreover, the elements can influence each other (Nkundabanyanga et al., 2014); for example, reduced opportunity may increase rationalization (Walters & Hunsicker-Walburn, 2015). Therefore, each security mechanisms should be examined systematically, and all possible interactions between the elements of the triangle should be thoroughly analyzed. To accomplish this task and identify the most effective combination of opportunity, rationalization, and need reduction mechanism that can substitute for proctoring, the taxonomy of cheating prevention techniques rooted in the fraud triangle theory is used.

The Taxonomy of Cheating Prevention Techniques

Educators and scholars who have been studied academic dishonesty among college students have developed numerous cheating prevention techniques (Moten et al., 2013; Shute & Rahimi, 2017; Srikanth & Asmatulu, 2014; Tinkelman, 2012). Several researchers organized these techniques with respect to their purposes (Hodgkinson, Curtis, MacAlister, & Farrell, 2016; Moten et al., 2013; Srikanth & Asmatulu, 2014; Stack, 2015), analyzed them through the lens of the fraud triangle theory (Lewellyn & Rodriquez, 2015; Tinkelman, 2012; Varble, 2014), and systematized in a taxonomy of cheating prevention mechanisms (Tinkelman, 2012; Varble, 2014). The taxonomy involves three categories: opportunity reduction, rationalization reduction, and need reduction (Tinkelman, 2012; Varble, 2014). The purpose of each category is to reduce, neutralize, or block the corresponding cheating behavior generated by perceived opportunity, need, and rationalization (Tinkelman, 2012; Varble, 2014). According to the fraud triangle theory, if any of the triangle's elements is reduced, neutralized, or blocked, less cheating should take place (Cressey, 1950; Becker et al., 2006; Lewellyn & Rodriquez, 2015; Tinkelman, 2012; Varble, 2014). The opportunity reduction category may involve time restriction, blocked backtracking (Beck, 2014; Stack, 2015; Varble, 2014), and higher order thinking levels of test items (Ladyshewsky, 2015; Varble, 2014). The rationalization reduction technique may include institutional policies, student honor codes (Corrigan-Gibbs et al., 2015; McCabe et al., 2012; Tinkelman, 2012), cheating statements on the syllabus (Varble, 2014), warning cheating statements before each test (Beck, 2014; Corrigan-Gibbs et al., 2015), and building an atmosphere of appreciative education where instructors and students respect each other (Hodkinson et al., 2016; Lang, 2013; Tinkelman, 2012). The need reduction category emphasizes the true value of education and acquired knowledge (Tinkelman, 2012; Ladyshewsky, 2015; Varble, 2014), importance of the course content for a future profession, assignments that involve students in active learning (Tinkelman, 2012). Although many of the reduction methods described in the literature can be used to prevent or reduce any form of academic dishonesty, the department where the study took place focused solely on nonbiometric mechanisms that have a potential to prevent or reduce cheating on web-based exams.

Web-based Exams' Opportunity Reduction Techniques

To control for the top opportunity risk factors, acquiring needed information from the Internet or a friend (Malgwi & Rakovski, 2008), different types of academic dishonesty associated with these risk factors should be considered (Hodgkinson et al., 2016; Moten et al., 2013; Srikanth & Asmatulu, 2014). During web-based exams students can copy from another student, in person or through electronic devices, use unauthorized materials such as the Internet, notes, solution manuals, study guides, collaborate with other students, and exhibit fraudulent behavior such as fake illness which prevent taking the exam during the designated time (Hodgkinson et al., 2016; Moten et al., 2013; Srikanth & Asmatulu, 2014). A simultaneous use of synchronous administering of exams (Moten et al., 2013; Stack 2015), strict time limit (Tinkelman, 2012; Varble, 2014), one attempt to take each test which should be completed once started (Moten et al., 2013; Varble, 2014), and randomization of questions and answer choices (Moten et al., 2013; Stack, 2015; Tinkelman, 2012; Varble, 2014) have a potential to significantly reduce or completely prevent opportunities to copy from another student, consult the textbook, or collaborate during the exam (Stack, 2015; Varble, 2014). Development of new exams each semester or multiple versions of the same exam during one semester prevents opportunity to acquire information about the exam questions and its solutions from students who took the exam previously (Tinkelman, 2012; Varble, 2014). A different version of an exam can be given to students who cannot take the exam at the designated time. Incorporation of essay questions or any unique questions focused on understanding, critical thinking, and interpretations may reduce opportunity to look for information on the Internet, in textbooks, study guides, or solution manuals (Ladyshewsky, 2015; Varble, 2014). Blocked backtracking option (Stack, 2015; Varble, 2014), when students cannot go back to the previous questions, and one question per page (Varble, 2014), may reduce opportunity to get help from a tutor or another person who does not take the exam during the same time (Beck 2014; Stack, 2015; Varble, 2014). Lockdown browser prevents accessing information from the Internet or from the computer and blocks copying the exam questions (Stack, 2015; Varble, 2014). Many opportunity reduction techniques described above were used in several studies that follow.

Studies with Emphasis on the Opportunity Reduction Techniques

Varble (2014), who compared 19 face-to-face students' scores on pencil-and-paper proctored exams with 28 online students' scores on secured unproctored exams in undergraduate marketing courses, focused on reducing opportunities to cheat during online exams based on the fraud triangle theory. The sections were offered in 2012 (D. Varble, personal communication, July 28, 2016) at a university in Indiana during the same semester and taught by the same instructor who used the same syllabus, textbook, assignments and exams in both classes (Varble, 2014). The unproctored exams were administered through Blackboard and incorporated the following security mechanisms: the questions were randomly selected from the publisher's test-bank, the question items and answers were shuffled, all exams had limited time, one attempt with the forced completion, one question visible at a time, blocked backtracking, and lockdown browser. Although Varble's (2014) taxonomy included rationalization and need reduction methods, his study focused on opportunity prevention mechanisms and incorporated only one rationalization reduction technique: the university student code of conduct was stated in the syllabus and discussed during the first week of classes in both sections of the course. The participants in Varble's (2014) investigation had to complete 20 multiple-choice questions within 25 minutes with an average of 1.25 minute per question. The exams questions, many of which were at remembering level of Bloom's (1964) taxonomy, tested concepts from the weekly reading assignments. There were 13 weekly exams and the final exam; the online students took all exams online, while face-to-face students took the first 11 weekly exams online, but the last two and the final exams in class in pencil-and-paper format (Varble, 2014).

Varble (2014) used a split-plot, or mixed, ANOVA and found that online students did significantly better on all 13 exams than face-to-face students (F (1, 45) = 3647.63, $p < 0.001$, $\eta^2 =.16$), with the biggest significant difference between proctored and unproctored exams (F (1, 45) = 17.25, $p < 0.001$, $d =1.24$). Similarly, on the final exam, the scores of face-to-face students ($t = 4.47$, $p < .001$, $d =1.38$). The researcher applied discriminant function analysis and found that remembering questions discriminated between proctored and online exams the most (*correlation discriminant* = .84, *standard discriminant* = 1.36) and analyze (*correlation discriminant* = .12, *standard discriminant* = -.41) and apply questions (*correlation discriminant* = .10, *standard discriminant* = -.54) the least. Varble concluded that, although the opportunity reduction mechanisms used in the study were not effective in decreasing the distinction between students' scores on proctored and unproctored exams to insignificant level, the use of fewer remember question and more understand, analyze and apply questions may reduce the distinction. It is harder to cheat when the answers cannot be looked up, but require using critical thinking skills and reasoning (Ladyshewsky, 2015; Tinkelman, 2012; Varble, 2014). The difference in the scores could be explained by different cheating factors.

Varble (2014) noticed that the fact that online students did significantly better on remember questions and less different on understand, analyze, and apply questions suggested that cheating through the use of prohibited sources could take place (Varble, 2014). All exams questions were based on the chapters' readings and provided by the publishers (Varble, 2014). The answers to these questions could be easily found on sites that help students with exams (Fisher, McLeod, Savage, & Simkin, 2015; Malesky et al., 2016) or in the textbook. Another factor can be related to asynchronous administration of the exams. The face-to-face students took the tests during the same class period. The online students completed the exams on the same day, but not synchronously with face-to-face students (D. Varble, personal communication, July 28, 2016). The online students could have significantly higher scores if the face-to-face students took the exams before the online students and shared some information about the exams with their online friends. The exams for the online students were open during the entire test day (D. Varble, personal communication, July 29, 2016). If some online students took the test earlier than other online students, the information about the exams' question could be shared.

Another study, which incorporated several techniques that suppress opportunities to cheat, was conducted by Beck (2014) in the spring 2012 semester at a university in Wisconsin with three sections of an undergraduate introductory economics course: online section with 19 students, hybrid section with 21 students, and face-to-face section with 60 students. Similar to Varble's (2014) study, the course assignments, exams, instructor, and semester were the same. The online students took the midterm and final web-based exams with the following security mechanisms: the questions were randomly chosen from the test bank available in D2L, the university LMS (V. Beck, personal communication, May 17, 2015), the exams had restricted time of 70 minutes for 50 multiple-choice questions with an average of 1.4 minutes per question, one attempt with the forced completion, one question visible at a time, and blocked backtracking (Beck, 2014). The online students had two days to take the exams, Thursday and Friday; they did not have access to the exams after the tests were completed. Unlike Varble (2014), Beck (2014) did not discuss different cognitive levels of exam questions and did not use lockdown browser. However, the researcher incorporated more rationalization prevention techniques than Varble (2014): before each exam, the students were warned not to cheat with the detailed explanation of the consequences of academic misconduct, and a message that the instructor could see what students were doing online was posted on the course website. The face-to-face students took the pencil-and-paper version of the same exams with scantrons in class on Thursday, while the hybrid students could take the same pencil-and -paper tests at the university proctoring center ether on Thursday or Friday. The proctors returned the pencil-and-paper exams with the scantrons to the instructor right after testing (Beck, 2014).

Beck (2014) conducted t-tests and found that there were no significant differences in students' scores on the proctored and unproctored midterm exam ($M_{proctored}$ = 37.75, *SD* =5.19, $M_{unproctored}$ = 38. 47, *SD* = 5.78, *t* = 0.534, *p* > .05) and on the final exam ($M_{proctored}$ = 40.21, *SD* = 4.72, $M_{unproctored}$ = 40.63, *SD* = 4.79, *t* = 0.347, *p* > .05). ANOVA indicated no significant difference in scores on both exams across delivery modes, face-to-face, hybrid, online, ($F_{midterm}$ = .0239, *p* =.788; F_{final} = .141, *p* =.869). Another ANOVA, which was conducted to compare students' academic abilities in face-to-face, hybrid, and online sections, showed no significant differences in GPA between the groups. Beck (2014) concluded that students' performance on unproctored exams might be comparable with the students' performance on proctored exams when the security mechanisms used in her study are incorporated. The difference in Varble (2014) and Beck's (2014) findings may be explained by the incorporation of the "cheating warning" rationalization technique.

Stack (2015) in his study with 287 online criminology university students in Michigan used almost the same opportunity cheating prevention techniques as Varble (2014) and Beck (2014), but made more emphasis on the importance of reducing the opportunity to disseminate exam questions before, during, and after an exam was administered. The researcher thoroughly explained the purpose of each security mechanisms used in the study. The Blackboard lockdown browser blocked all other computer’s functions that prevented emailing, web browsing, and copying the exam questions (Stack, 2015). The exam items were scrambled, which reduced opportunities to collaborate if students completed the exam side-by-side. All participants took the unproctored exam synchronously, which eliminated the opportunity to take the exam at one time and help a friend to take the same exam at different time (Stack, 2015). This feature, combined with the limited time, also did not allow for dissemination of the exam’s content among the students while the exam window was open. The students could not access the exam before and after scheduled time; blocked backtracking and time restriction, which allowed for one minute answering each question, further reduced opportunity to cheat (Stack, 2015). The researcher did not discuss the levels of Bloom’s taxonomy of the exam questions and did not list the incorporation of exam questions that require higher order levels of thinking as a security mechanism. Stack (2015) did not specify whether any rationalization or need prevention mechanisms were used.

Similar to Varble (2014) and Beck (2014), Stack (2015) controlled for instructor and course content effects: all sections were taught by the same instructor, who used the same textbook, assignments, and exams. However, unlike Varble (2014) and Beck (2014), Stack (2015) did not control for the history effect: he used the scores of students in 10 online criminology courses offered during Fall 2006 through Fall 2011. The five courses offered at the beginning of the study had the proctored pencil-and-paper multiple-choice final exams with scantrons, while the other five courses had an equivalent unproctored web-based version of the same final exam (Stack, 2015).The courses, two sections of which were offered every semester, were not randomly assigned to treatments (S. Stack, personal communication, July 5, 2015).

The regression analysis showed that there was no trending in exams' scores during the study's period (Stack, 2015). Moreover, no significant difference in student performance on proctored and unproctored final exam was found (R^2 =. 343, b = 1.08, p > .05). The researcher concluded that the combination of the cheating reduction techniques used in his investigation might have a security level comparable with proctoring (Stack, 2015).

Ladyshewsky (2015) implemented web-based exam security mechanisms in his study with a convenient sample of 250 postgraduate management and leadership business students during the 2012 through 2013 (R. Ladyshewsky, personal communication, July 29, 2016) academic year in Australia. The participants were enrolled in nine face-to-face sections of the course: some of these sections were offered in a blended-intensive mode during the day two or three times a week during two or three weeks, while others were offered once a week in the evening over the period of 14 weeks (Ladyshewsky, 2015). The course content, assignments, and exams were the same across all sections; however, unlike Varble (2014), Beck (2014), and Stack's (2015) investigations, different instructors taught three out of the nine sections in Ladyshewsky's (2015) study. The higher levels of Bloom's taxonomy understand, analyze, and apply exam questions were the main opportunity reduction technique used by the researcher. Randomization, restricted time, and blocked backtracking were other opportunity prevention mechanisms incorporated in the study. The study also incorporated one rationalization prevention mechanism: the university where the investigation took place had a very strict enforced academic integrity policy, which could significantly deter students from cheating (Ladyshewsky, 2015). The lockdown browser was not used by Ladyshewsky (2015).

A purposely developed by the researcher multiple-choice pencil-and-paper test included 49 short-case scenario questions that required critical thinking and one noncritical thinking question that could be easily answered by all students (Ladyshewsky, 2015). This pencil-and-paper test was supervised and administered to 136 students during the first four trimesters. The students had 130 minutes to answer 50 questions, on average 2.6 minutes per question. The tests and answer sheets were collected immediately after the exams to prevent circulation of the questions among the students (Ladyshewsky, 2015). The pool of the 50 questions of the pencil-and-paper test was created in Blackboard. Twenty-five multiple-choice items were randomly drawn from this pool of 50 questions for the online exams administered to 114 students during the next five terms. The participant had a four-day period to complete this 70-minute online exam, on average with three minutes per question (Ladyshewsky, 2015).

Ladyshewsky applied a one-way ANOVA for all nine tests and found a significant difference in scores ($Mp_1 = 71.8$, $Mp_2 =71.3$, $Mp_3 =76.1$, $Mp_4 =78.3$, $Mup_5 = 71.0$, $Mup_6 = 68.3$, $Mup_7 = 73.8$, $Mup_8 = 72.4$, $Mup_9 = 67.4$; F (8, 241) = 3.628, $p = .001$). The second, sixth, and ninth term courses were taught by different instructors. On average, while there was no pattern in unproctored exams, the scores demonstrated an increasing pattern in proctored exams (Ladyshewsky, 2015), which could be an evidence of dissemination of exam questions with time. The lowest scores were in the sections taught by the different instructors; thus, the difference in scores could occur because the instructor who was teaching the course on a regular basis could teach more towards the test since it was developed, which resulted in higher average scores (Ladyshewsky, 2015). However, ANOVA for the six sections taught by the same instructor also indicated a significant difference in scores (*F* (5, 155) = 2.612, *p* = .027), but not toward more cheating on unproctored exams (*Mp* = 75.4, *Mup* = 72.4). Although the means of scores on proctored and proctored exams were different, it is not clear whether this difference was significant: the researcher did not use any inferential statistical techniques to test the significance of the difference because he was mostly looking at what occurred over time and whether there was increasing behavior in scores (R. Ladyshewsky, personal communication, July 29, 2016). However, he gave possible explanations of why the scores' on the unproctored exams were lower.

Ladyshewsky (2015) suggested that the students performed worse on unproctored exams due to the fact that the web-based exams could be harder than proctored exams because the randomization of questions on web-based exams could result in the bigger proportion of harder questions. Moreover, the exam questions required higher order thinking skills applied to a business scenario; the answers to these questions could not be found on the Internet and were hard to memorize. Most of the students who took unproctored exams were adults in their mid-thirties enrolled in the blended intense sections of the course due to their extremely busy schedule (Ladyshewsky, 2015). These students could have less time to study than traditional students who took proctored exam (Ladyshewsky, 2015) and could cheat less because research suggested that older students are less likely to cheat (Ladyshewsky, 2015; Prince, Fulton, & Garsombke, 2009). The study's online tests were designed in accordance with the best practices described in the literature, which could also reduce propensity to cheat (Ladyshewsky, 2015). Because the mean of scores on the unproctored exams was lower than on proctored, Ladyshewsky (2015) concluded that the higher order thinking exam questions in a combination with randomization and blocked backtracking were effective cheating reduction mechanisms. A thorough analysis of the opportunity reduction mechanisms used by the department where the given investigation took place is provided below.

Application of the Opportunity Reduction Techniques to the Study

Synchronous administering of the unproctored web-based exam is the first opportunity reduction technique utilized by the department. The students in all web-based introductory statistics sections take Set 1 and Set 2 unproctored web-based exams on the same day during the same time frame. The purpose of this opportunity reduction mechanism is to eliminate the possibility for one student taking an exam at one time and then helping a classmate with the same exam at another time (Stack, 2015). This technique, in combination with the restricted time, also prevents dissemination of exam items while the exam window is still open (Moten et al., 2013; Stack, 2015). Additionally, when synchronous testing is used, instructors do not need to create a new version of the test for each section because all ten sections take the exam at the same time. However, synchronous testing may trigger rationalization that students did not know that the quiz was not open during the whole day and did not allocate enough time to study, and may increase opportunity for collusion when two or more students work on the same exam side-by-side. To neutralize the first side effect, the dates and times of all unproctored exams are announced and posted on the course web-site on the first day of classes so that students can do all necessary adjustments in their schedules in advance. To neutralize collusion, randomization of questions and multiple-choice answers, one question per page, and blocked backtracking are used.

In Varble's (2014) study, the results of which showed significantly higher scores on unproctored exams, the online exams were open during one day. The asynchronous testing could allow for dissemination of test items (Moten et al., 2013; Tinkelman, 2012), which could be one of the reasons why students performed better on unproctored exam. The students in Beck's (2014) study, in which a significant difference in students' scores was not found, could take the unproctored exam throughout two days. However, the researcher used warning cheating messages (Beck, 2014), which could neutralize cheating opportunity of sharing exam questions (Corrigan-Gibbs et al., 2015). Moreover, unlike the department where the given investigation took place, in which up to 10 sections of the course take unproctored exams, Beck (2014) had only one small group of online students (*N* =19); massive circulation of test items among the participants was, perhaps, not expected. In Ladyshewsky's (2015) study, the unproctored exams were open throughout four days, but the researcher did not find any evidence of cheating because the students did worse on the unproctored exams. However, dissimilar to mature graduate working students in their mid-thirties, who tend to cheat less (Ladyshewsky, 2015; Prince et al., 2009; Siniver, 2013), younger undergraduate and community college students may take advantage of asynchronous testing and share the tests' items with their friends (Fask, 2015; Moten et al., 2013; Srikanth & Asmatulu, 2014). Unlike Varble (2014), Beck (2014), and Ladyshewsky (2015), Stack (2015) utilized synchronous testing and found it to be effective in suppression of cheating. Therefore, synchronous testing selected by the department as a main reduction opportunity mechanism has a high potential to neutralize cheating.

Limited test taking time is the second opportunity reduction technique used by the department. The purpose of this technique is to restrict the testing time to the period sufficient to answer exam questions but not sufficient enough to look up the answers on the Internet, in the book, solution manuals, and study guides, or text and call friends (Bain, 2015; Hodgkinson et al., 2016; Moten et al., 2013; Srikanth & Asmatulu, 2014). Restricted time also makes it difficult for students to help each other during collusions (Moten et al., 2013; Stack, 2015). However, limited test taking time may increase the rationalization risk factor that the test is too hard and the need risk factors related to the fear of getting bad grades (Harding, 2001; Hodgkinson, 2016; Malgwi & Rakovski, 2008). To neutralize the rationalization that the exam is too hard, the sufficient test time should be carefully identified such that an average student can answer all exam questions during provided period without rushing (Harding, 2001; Hodkinson, 2016). The math department's experience in administering online exams at the college where the given study took place and the literature were used to determine the sufficient testing time for the web-exams involved in the study.

The math department has been administering web-based exams in online and hybrid sections since 2011. The average time taken by the students for each exam in these sections was analyzed. The results showed that about 70 minutes are needed for a secured, web-based exam with 23 questions. This result is consistent with Harding's (2001) recommendations of giving students 3-4 times longer period to take the test than the instructor would need. Three instructors of the department completed the study's secured, web-based exam independently with the mean testing time of 17 minutes; 17(4) = $68 \approx 70$. Seventy minutes for 23 questions yield a time frame of three minutes per question.

Varble's (2014) students had on average 1.25 minutes to answer each exam question, Beck (2014) allocated 1.4 minutes for this purpose, and Stack's (2015) participants had only one minute per question. Unlike these researchers, Ladyshewsky (2015) incorporated high levels of Bloom's (1964) taxonomy questions and allocated three minutes to answer each of these questions. The given study also incorporates exam questions that require higher order thinking. Therefore, three minutes per question is a sufficient period for students to complete the test without rushing.

To further reduce the rationalization that the exam is too hard and neutralize the need of fear of bad grades, the participants of the study were given a web-based practice test before each exam, the structure and time frame of which were identical to the actual exams. During these practice tests, the students became familiar with the test's procedure and learned how to manage their test time accordingly. The acquired test-taking skills during the practice tests may reduce fear of bad grades (MacGregor & Stuebs, 2012; Tinkelman, 2012) and concern that time restriction makes the test harder.

Randomization of exam questions and multiple-choice responses, a feature available in the college's LMS, is another opportunity reduction technique used by the department. If two or more participants sit next to each other, it is highly likely that each of them can see different questions (Beck, 2014; Moten et al., 2013; Hodgkinson, 2016). Unlike Ladyshewsky's (2015) study, in which a subset of questions was randomly selected from the created pull possibly generating nonequivalent exams, each student in the given investigation answered the same set of questions, but in a different order. Thus, the randomization of questions in the given study preserves equivalency of exams. Several researchers found that randomization of high-quality exam questions does not impact students' performance (McLeod, Zhang, & Yu, 2003; Tal, Akers, & Hodge, 2008; Xu, Kauer, & Tupy, 2016). Therefore, to prevent rationalization that the exam is not fair and need to cheat generated by fear of getting bad grades in relation to randomization of exam questions, the exam items should be developed in accordance with the best practices of test questions' creation (Tinkelman, 2012; Xu et al., 2016).

The department also incorporated one exam question per page and blocked backtracking. The purpose of both opportunity reduction mechanisms, in combination with the restricted testing time, is to reduce collaboration during exams (Beck, 2014; Stack, 2015; Tinkelman, 2012). If students could see more than one question at a time and could go back, they could ask a friend to solve a problem while they were working on another one (Beck, 2014; Stack, 2015). One question at a time and blocked backtracking may trigger rationalization that the exam is not fair for online students because students in face-to-face classes can work on pencil-and-paper tests' questions in any order (Ladyshewsky, 2015). The detailed explanations that each secured, web-based exam is a part of the curriculum and administered in all sections offered by the department regardless of the delivery mode may neutralize this rationalization (Malgwi & Rakovski, 2008). Clear course expectations, discussions about the goals' of the department focused on credibility of the offered courses, high standards, and effective learning and instruction may further decrease the concern about "unfairness" of backtracking (Hodgkinson, 2016; MacGregor & Stuebs, 2012; Tinkelman, 2012).

Deferred feedback is another opportunity prevention technique, the purpose of which is to prevent distribution of solutions and answers to exams' questions (Beck, 2014; Tinkelman, 2012). When students answered a question, the correct answer, hints, and marking score were not provided; thus, students did not know whether they answered a question correctly or not while the exam window was still open. Because up to 11 sections of the course were involved each semester, to avoid circulation of exam items among the students, the instructors decided to not open both sets of exams for students' review at all. However, to reduce rationalization that the exam is not fair and fear of getting undesired grades, and inform the students about their exams' errors, after each exam was over, each student received individual feedback on every incorrect answer and conceptual misunderstanding.

Similar to Ladyshewsky's (2015) study, the use of higher order thinking exam questions is an important opportunity reduction technique utilized by the department. All study's exams included about 80% of questions, answering of which requires statistical reasoning, critical thinking, and interpretation. The nature of these questions itself precludes cheating because the answers to them cannot be looked up online or in printed resources (Ladyshewsky, 2015; Varble, 2014; Xu, Kauer, & Tupy, 2016; Zito & McQuillan, 2010). This opportunity reduction mechanism may increase rationalization that the exam is too hard and need generated by fear of getting undesired grades (Malgwi & Rakovski, 2008). However, in addition to the study guides, practice tests, and review sessions that reduce the rationalization and need described above (Tinkelman, 2010; Varble, 2014), the entire course curriculum, which is based on inquiry learning, develops statistical reasoning and critical thinking, preparing students for answering higher order thinking questions successfully (Jensen, McDaniel, Woodard, & Kummer, 2014). Discussions about value of knowledge, understanding, and logical reasoning may inspire students to study harder and be better prepared for exams, which automatically reduces fear of tests and fear of undesired grades (MacGregor & Stuebs, 2012; Tinkelman, 2012).

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 are other opportunity reduction techniques used by the department. Both mechanisms were used to decrease circulation of exam items among the students (Tinkelman, 2012; Stack, 2015). Beck (2014) and Stack (2015) utilized lockdown browser for this purpose and to reduce opportunities looking up the needed information on the Internet. However, lockdown browser was not suitable for the department because of three reasons.

First, this security mechanism was not available at the college where the study took place. College students cannot afford lockdown browser's individual licenses due to the high cost. Second, 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. About 97% of college students have portable devices that they carry around on a regular basis (Walters & Hunsicker-Walburn, 2015). Access to the Interne, blocking of which is the main purpose of lockdown browser (Stack, 2015), may become available on these devices instantaneously just with one click (Khan & Balasubramanian, 2012; Walters & Hunsicker-Walburn, 2015). Lastly, the Internet is not very useful when higher order thinking exam questions uniquely created by faculty are incorporated (Ladyshewsky, 2015; Varble, 2014): even if students go on the Internet during the exams, they are not able to find the answers there. Because each opportunity prevention technique should be used when it is highly needed due to the fact that it may increase rationalization and trigger more cheating (MacGregor & Stuebs, 2012; Walters & Hunsicker-Walburn, 2015), lockdown browser was not used by the department. The provided analysis suggests that the opportunity prevention mechanisms selected based on the criteria described above have a high potential to minimize opportunities to cheat to the level comparable with the level of cheating opportunities during proctored exams.

I organized the described above opportunity reduction techniques with their purposes, possible influence on other triangle’s factors, and neutralization of the negative side-effects in a table provided in Appendix A. This table, which is a concise description of the opportunity reduction factors used by the department, can also be used as a checklist in identifying a web-based exam’s security strength. Although rationalization was mentioned several times as a factor which can be triggered by opportunity reduction techniques, there are techniques that can specifically reduce student rationalization to cheat.

Web-based Exams' Rationalization Reduction Techniques and Related Research

The fraud triangle's rationalization factor can be divided into four major categories: claiming ignorance due to ambiguous policies and instructions or their absence, claiming unrealistic expectations of the instructor, looking at peer behavior, and claiming that the fraud is minor and does not matter (MacGregor & Stuebs, 2012). Although rationalization is harder to control than cheating opportunity (Becker et al., 2006; Malgwi & Rakovski, 2008), instructors have some influence on this factor and can positively affect students' attitude by addressing rationalization before cheating occurs (Hodgkinson, 2016; Jones, Blankenship, & Hollier, 2013; Tinkelman, 2012; Varble 2014).

Rationalization *I did not know it was cheating*, a frequently used justification of academic misconduct (Tinkelman, 2012; Varble, 2014), corresponds to ignorance due to ambiguous or not clearly stated policies and reflects an attitude that any action is acceptable until it is not specifically prohibited (MacGregor & Stuebs, 2012). Unclearly stated policies about the use of printed and electronic resources during online exams can create ambiguity (Jones et al., 2013; MacGregor & Stuebs, 2012). Jones et al. (2013) conducted a stud with 194 paralegal and business online university students and found that over 60 % of them used open books and or notes during an unsupervised web-based exam. About 58% of the participants reported that they believed that the use of these resources did not constitute cheating on online tests (Jones' et al., 2013). King, Guyette, and Piotrowski (2009) investigated rationalizations of cheating behaviors during online exams in their study with 121 undergraduate business students in a university in the South. The researchers used a Likert-type scale to measure the participants' rationalizations in two conditions: (a) when no policy on test taking was given and (b) when the cheating behaviors during online testing were clearly identified by the instructor (King et al., 2009). The results of the study showed that after what constitute cheating behavior during online tests was explained to the students, their perceptions that using printed materials during online exams is cheating increased from 7% to 71%; consulting with other individuals during exams increased from 50 % to 82%, using personal class notes increased from 9% to 68%, and utilizing online sources increases from 21% to 78% (King et al., 2009). Thus, rationalization of not knowing what is cheating may be prevented by including clear descriptions of what constitute academic misconduct in the class syllabus, cheating warning statement, and class discussions (MacGregor & Stuebs, 2012; Tinkelman, 2012; Varble 2014).

Honor code is another action that can reduce rationalization of not knowing what is cheating (Tinkelman, 2012; Varble, 2014). McCabe et al. (2012) found that students at colleges with honor code tend to cheat less than students from campuses where honor code was not adopted. However, Corrigan-Gibbs et al. (2015) found that cheating warning messages are more effective than honor code. The researchers randomly divided 409 participants into three groups: in one group the honor code was used, in the second one a warning message, and the third group was a control group without honor code or warning message. Corrigan-Gibbs et al. (2015) created a web-site where they posted complete answers to the free response questions of the online exam. About 25% of the participants cheated using the posted answers. Only 7% of the cheaters were in the cheating warning messages group; others were in the honor code and control group (Corrigan-Gibbs et al., 2015).

Rationalizations *the course or test is too hard, the instructor is a hard grader, students who work hard deserve a good grade*, which can be classified as claims of unrealistic expectations of the instructor, are the top rationalization risk factors identified by students (Malgwi & Rakovski, 2008). Out of all rationalizations, faculty have influence on these factor the most because they are related to the quality of teaching (Jones' et al., 2013; MacGregor & Stuebs, 2012). Well-developed curriculum with real-world applications relevant to students' goals, clearly stated course expectations, fair grading based on well-designed grading rubrics, frequent feedback, and reasonable workload are some effective prevention rationalization techniques (Hodgkinson, 2016; Tinkelman, 2012; Varble, 2014). Additionally, fairness of the used tests should be considered (Harding, 2001; Hodgkinson, 2016; Tinkelman, 2012).

Harding (2001) in his study with 65 university engineering students examined their perception about what instructors can do to reduce cheating on exams. The researcher found that the majority of the participants' responses (4.45 out of 5.00) indicated that the students would be less likely to cheat if instructors' tests would be fair: challenging, but doable, with enough time to complete the task, and not focused on pointless memorization. The list of the concepts covered on the test and previous exams posted on the course web-site, reference sheet with formulas allowed during exams, high quality study guides, practice tests with feedback, and exams' study sessions are other mechanisms that may neutralize the top rationalization risk factors (Harding, 2001; Hodgkinson, 2016; Tinkelman, 2012). If clear description of topics covered on each test and old exams are available to students, they will not be able to say that exams' expectations were not clear to justify possible cheating (Tinkelman, 2012). If several instructors teach the same course, consistency in pedagogy and content delivery, administering the exams, and grading, may reduce rationalizations about difficult exams and hard graders (Malgwi & Rakovski, 2008). Providing in advance clear explanations of what materials are covered on the exams and making previous year exams available may also reduce perceived opportunity to cheat: there is no need asking friends in prior classes about old exams' content and answers if the old exams are available to all students (Tinkelman, 2012).

Students' claims that everyone is cheating and nobody got hurt belong to the last two rationalization categories: influence of peer cheating behavior and claiming that the fraud is minor and does not matter. The literature suggests that detailed descriptions of academic misconduct's consequences, faculty intolerance to cheating, emphasis on the importance of acquired knowledge and ethical behavior, discussions on how cheating negatively impact other students in the class are techniques that can reduce these two rationalizations (Hodgkinson, 2016; MacGregor & Stuebs, 2012; Tinkelman, 2012). Some studies showed that all rationalizations can be reduced or neutralized by an atmosphere of mutual respect and trust when instructors care about student learning (Harding, 2001; MacGregor & Stuebs, 2012).

The Study's Rationalization Reduction Techniques

The instructors of the course developed common syllabus with clearly stated cheating policies and consequences of academic dishonesty. The purpose of this reduction mechanism is to prevent rationalization of not knowing what is cheating (Beck, 2014; Tinkelman, 2012; Varble, 2014). Before each unproctored exam, the students were told that the instructor can see some of their online actions and were warned not to cheat: the warning statement was emailed to the class and posted on the course website. This prevention technique reduces rationalizations *it was not clear what we could not use on this exam, the instructor does not care about cheating and does not try to prevent it*, and *there are no severe consequences* (Beck, 2014; Tinkelman, 2012). Noting that the instructor can see some students' actions can reduce cheating opportunity factor (Beck, 2014). Honor code, another rationalization of not knowing what is cheating prevention technique (Tinkelman, 2012), was not used by the department because the college where the study took place does not have it, and because warning messages were found to be more effective than honor code (Corrigan-Gibbs et al., 2015). To reduce rationalization that the course is too hard (Tinkelman, 2012), the course syllabus had clearly identified objectives and course expectations. The list of all concepts covered by each exam, old exams with solutions available to students during the entire semester (Tinkelman, 2012), study sessions before each exam, and practice tests, the structure, covered topics, and security mechanism of which are identical to the actual exam are techniques that neutralize rationalizations that the exam is too difficult and the exams expectations are unclear. The practice tests were opened one week before the scheduled exam and became invisible to students during the actual exam to prevent opportunities to look at the practice tests' content during the exam. All face-to-face and hybrid sections had in class exams' study sessions, the video recordings of which were posted on the course websites for all delivery modes, including fully-online sections. These unclear exam expectations rationalization prevention mechanisms may neutralize fear of undesired grade need factor (Malgwi & Rakovski, 2008; Tinkelman, 2012).

The department where the given study took place thoroughly designed web-based exams in accordance with the best test creation practices. Fair automatic grading based on questions' scores inserted in the LMS (Tinkelman, 2012), carefully identified sufficient test taking time (Harding, 2001), and the list of all needed formulas provided on each exam reduce unfairness of tests rationalization factors (Malgwi & Rakovski, 2008; Tinkelman, 2012). The list of formulas also prevents the perceived cheating opportunity factor looking them up on the Internet or in printed resources (Tinkelman, 2012). Well-developed curriculum with real-world applications relevant to students' interests reduces rationalization risk factor related to inapplicability of the covered material. This mechanism may reduce the fear of undesired grades need because students are willing to study harder the material that is important to them (Tinkelman, 2012). Better exams preparations reduce the fear of undesired grades (Harding, 2001; Tinkelman, 2012). The department's emphasis on the importance of acquired knowledge and ethical behavior and focus on building atmosphere of mutual respect and trust between instructors and students may reduce all rationalization and neutralize need factors (Harding, 2001; Lang, 2013; MacGregor & Stuebs, 2012; Tinkelman, 2012). Students would cheat less when they recognize the importance of knowledge (Lang, 2013; Tinkelman; 2012) and feel the instructor's care and respect (Lang, 2013; MacGregor & Stuebs, 2012).

The rationalization prevention mechanisms described above have a high potential to minimize rationalization to cheat, which further secure unproctored web-based exams. Similar to the opportunity reduction techniques, I organized the rationalization reduction mechanisms, their purposes, possible influence on other triangle's factors, and neutralization of the negative side-effects, if any, in a table provided in Appendix B. The last fraud triangle element need was also considered by the department.

Need Reduction Techniques and their Use in the Study

Students' needs or incentives to cheat usually have economic or social nature (MacGregor & Stuebs, 2012). Fear of getting undesired grades is related to economic incentive: students are afraid of losing financial aid, opportunity to be accepted to good universities, and get high-payed jobs (MacGregor & Stuebs, 2012; Tinkelman, 2012). Social needs are triggered by relationships with people: peer or parents pressures (MacGregor & Stuebs, 2012). Faculty have little control of need factor (Becker et al., 2006; Malgwi & Rakovski, 2008), but can influence this element of the fraud triangle indirectly through opportunity and rationalization prevention techniques (Tinkelman, 2012). If students know the concepts covered on each exam, obtain timely feedback from their instructors, and have access to other preparatory resources, their fear of getting undesired grades will decrease, and they will not have a need to cheat (Tinkelman, 2012).

Peer and parent pressure might be partially neutralized by discussions on how academic dishonesty affects other students and building an atmosphere of mutual respect in the classroom (Lang, 2013; Tinkelman, 2012). All these approaches are used by the department.

Other Theories and the Choice of the Study's Framework

The theory of planned behavior (Ajzen 2002; Imran & Nordin, 2013), goal orientation theory (Alt & Geiger, 2012; Zito & McQuillan, 2010), and item response theory (Champlain, 2010; Templin, 2016) are some other theories that can explain academic cheating. According to the theory of planned behavior, intention to participate in academic misconduct can be explained by attitude towards behavior, subjective norms, and perceived behavioral control (Ajzen 2002; Imran & Nordin, 2013). Attitude towards behavior represents individual's perception of a behavioral actions and their impact, subjective norms reflect normative assumptions of other people about cheating; perceived behavioral control explains whether an action or behavior is difficult or easy to perform (Imran & Nordin, 2013). The theory of planned behavior, which describes psychological reasons of why students can cheat, was not suitable for the given study because the study did not investigate reasons for cheating.

Goal orientation theory suggests that students' motivation for learning can be goal oriented or performance oriented (Alt & Geiger, 2012; Cheung et al., 2016; Zito & McQuillan, 2010). Goal oriented people truly value knowledge and focus on learning and understanding of the subject matter regardless of the fact whether an assignment is graded or not (Alt & Geiger, 2012; Zito & McQuillan, 2010). Performance oriented individuals do not want to spend their time and effort learning something that is not graded (Zito & McQuillan, 2010). According to the goal orientation theory, cheating reflects students' perceptions of the goal of an exam and what they gain by taking it (Zito & McQuillan, 2010). Students whose motivation is performance oriented tend to cheat more (Alt & Geiger, 2012; Zito & McQuillan, 2010). The goal orientation theory, which describes motivation behind academic dishonesty (Alt & Geiger, 2012; Cheung et al., 2016; Zito & McQuillan, 2010), was not chosen because the given study did not aim to investigate the motivation for cheating.

The item response theory is rooted in the computational analysis of students' responses (Champlain, 2010; Shu et al., 2013; Wollack, Cohen, & Eckerly, 2015). The computational analysis compares the actual test scores and estimated performance based on students' abilities, taking into consideration the source of error that occurs due to unobserved abilities (Champlain, 2010). The item response theory, a framework for detecting cheating (Shu et al., 2013; Wollack et al., 2015), was not chosen because cheating detection was not the focus of the study.

The given study was designed to determine whether systematically selected cheating prevention mechanisms can substitute for proctoring. This purpose drove the choice of the study's theoretical framework. The taxonomy of cheating prevention mechanisms (Tinkelman, 2012; Varble, 2014) rooted in the fraud triangle theory (Cressey, 1950) was the most appropriate theoretical foundation for the given study because it explains which security mechanisms have the potential to reduce factors needed for cheating to take place (Cressey, 1950; Tinkelman, 2012; Varble, 2014). If the factors needed for cheating are minimized, student performance on proctored and unproctored exams may be similar and proctoring may not be necessary.

Literature Review Related to Key Variables and Concepts

In the given quantitative study, which utilized a quasi-experimental one-group sequential design, I investigated whether there is a relationship between the web-based exam format, proctored versus unproctored, and student exam scores. This question was answered by testing the hypothesis that there is no significant difference in student performance on proctored and unproctored equivalent web-based exams when the same security mechanisms are used. Thus, the study's constructs of interest were web-based testing, challenges associated with cheating, and security mechanisms. The study's main independent variable was the format in which the exams were administered, proctored versus unproctored (IV1), and the dependent variable was the web-exam score (DV). The additional independent variables were the effects of the order in which the proctored and unproctored exams were administered (IV2), course delivery modes (a) web-assisted face-to-face, (b) hybrid, and (c) fully-online (IV3), and the instructor of the course (IV4). This section of the chapter describes the literature review of the study's constructs of interest, chosen methodology and methods, and justification of the rationale for the selection of the study's variables. The analysis of prior and current studies related to the independent and dependent variables and the research question of the investigation conclude the section.

Web-based Assessment and its Evolution

Emerging computing, information, and communication technologies have been deeply influencing academic assessment since the invention of microcomputers in the 1970s (Bennett, 2008; Schlegel & Gilliland, 2007). The first primitive computerized tests were introduced in the mid-1970s (Mazzeo & Harvey, 1988) and reflected the same properties as corresponding true-false pencil-and paper tests (Bennett, 2008; Bunderson, Inouye, & Olsen, 1989): the tests items were presented in a fixed order, the examinee were allowed to review, revise, and skip a test item (Mazzeo & Harvey, 1988; Patelis, 2000). Because all students were taking exactly the same exam, the first computerized tests had low security (Patelis, 2000). One of the first true-false computerized tests was the Minnesota Multiphasic Personality Inventory test, in which one item at a time was presented, and students responded by inserting “t” for true, “f” for false, and “?” for cannot say (Lushene, O'Neil, & Dunn, 1974). This generation of tests was administered on stand-alone computers as a one-time assessment event (Bennett, 2008; Bunderson et al., 1989) and was more expensive than the corresponding paper-and-pencil version (Bennett, 2008). Further evolution of test algorithms rooted in classical test theory (Bunderson et al., 1989) and innovations in hardware and software introduced automatically graded multiple-choice, fill-in-the-blank, short answer, and matching test items (Bennett, 2008).

The development of the item-response theory, which allowed for tailoring the tests’ difficulty and contents based on the examinee’s answers, resulted in the appearance of computerized adaptive tests in the mid-1980s (Bennet, 2008; Bunderson et al., 1989).

At that time, computerized adaptive assessments were mainly used in large-scale placement testing provided by testing companies at the designated testing centers and served mostly institutional needs (Bennett, 2008; Redecker & Johannessen, 2013). One of the first adaptive tests, the College Board Computerized Placement test, now known as Accuplacer, was implemented by the Educational Testing Service in 1986 (Bennet, 2008). The adaptive tests introduced blocked backtracking approach which does not allow for reviewing, revising, and skipping the test item. Blocked backtracking, together with the restricted exposure to all test questions, increased security of computerized testing (Patelis, 2000). Moreover, blocked backtracking allowed for identifying the exact number of minutes spent by a student on each exam item (Patelis, 2000).

The next generation of communication and information technology reflected a significant qualitative improvement in test administration and changed the nature of assessment (Bennet, 2008; Redecker & Johannessen, 2013; Timmis et al., 2015).The development of the Internet and dissemination of computers in daily life created favorable conditions for diffusion of all forms of computer-based testing in educational institutions (Dahalan & Hussain, 2010; Redecker & Johannessen, 2013; Timmis et al., 2015). Large-scale computerized-tests administered at special testing centers and traditional in-class pencil-and-paper tests were superseded by more flexible and cost-efficient web-based exams, which could be taken on computers, laptops, iPads, tablets, cell-phones and other similar electronic devices at any location with internet access (Redecker & Johannessen, 2013; Timmis et al., 2015). The era of web-based exams had begun.

Colleges and universities integrated administration of computerized placement tests provided by testing companies and publishers on their campuses (Bennet, 2008). Currently, the previously mentioned Accuplacer assessment provides over eight million diagnostic and placement tests and serves more than 2.5 million students a year on 1500 university and colleges campuses (College Board, 2016). The use of computer-based annual state assessment in schools has increased across the country (Blazer, 2010: Smarter Balanced Assessment Consortium, n.d.). Starting spring 2014, 17 states in the US, including California where the given study took place, have been administering end-of-the year web-based Smarter Balanced Assessment in English and mathematics instead of the previous standardized pencil-and-paper test. In spring 2015, this test was taken by more than six million students in grades 3-11, making it one of the biggest online assessments (Smarter Balanced Assessment Consortium, n.d.). Widespread adoption of LMSs with embedded test-building programs brought web-based exams into online, hybrid, and face-to-face classrooms (Arnold, 2016; Dahalan & Hussain, 2010; Ladyshewsky, 2015; Shute & Rahimi, 2017). Flexibility in creating online tests, building them from scratch, or quickly assembling them through the use of test-banks provided by publishers, automatic scoring and automatic recording of scores in the gradebook, basic test items' analysis, immediate feedback, and numerous inbuilt security mechanisms has made web-based testing popular among faculty (Dahalan & Hussain, 2010; Hameed, 2016; Ladyshewsky, 2015). Multimedia, simulation, game-based and other interactive test-items features brought a new potential for increasing student engagement and performance (New Media Consortium, 2016; Redecker & Johannessen, 2013; Timmis et al., 2015). Web-based exams have begun serving not only institutional, but also individual faculty and their students' purposes (Dahalan & Hussain, 2010; Redecker & Johannessen, 2013; Timmis et al., 2015). The advantages associated with web-based testing described above introduced new opportunities for teaching and learning which were reflected in the research literature discussed below.

Studies Related to Advantages of Web-based Exams

A number of studies showed that, since the early stages of LMSs' adoption in the beginning of the 21 century, constantly evolving LMSs and their inbuilt assessment features have been positively influencing teaching and learning (Daniel & Broida, 2004; Hanson & Robson, 2004; Lonn, & Teasley, 2009). In 2002-2003, an association of IT professionals EDUCAUSA conducted a pilot case study at three private American universities with highly selected enrollment, Brandeis University, Wesleyan University, and Williams College, to examine which LMSs' features were beneficial for faculty and students (Hanson & Robson, 2004). At that time, Williams used Blackboard LMS, Brandies utilized WebCT, and both Blackboard and WebCT were implemented at Wesleyan (Hanson & Robson, 2004). Blackboard and WebCT, created in 1997-1998 (Dahlstrom, Brooks, Bichsel et al., 2014), were the two most popular platforms in the US higher education at the early 2000s (Hanson & Robson, 2004). The researchers Hanson and Robson (2004) measured perceived LMSs' benefits through survey and focus groups with 981 students and 341 instructors. The study's findings showed that only about 43% of participating faculty utilized LMSs for teaching. These instructors used LMSs mostly for disseminating class information and materials; online quizzes and exams were used only by a few instructors. However, students whose instructors utilized online assessment reported that they found online quizzes/exams, immediate access to exam grades, and instructors' feedback highly beneficial for their learning (Hanson & Robson, 2004). The instructors who did not use the LMSs and their web-based assessment features explained that they were not trained how to utilize web-based testing and it was not clear for them whether this form of assessment could be more beneficial for teachers and students than old-fashion pencil-and-paper tests. They were also concerned about possible academic fraud (Hanson & Robson, 2004). Hanson and Robson (2004) suggested that further studies should investigate the influence of web-based exams on instruction and student performance.

Daniel and Broida's (2004) quantitative study was focused on the investigation of a possible positive impact of web-based assessment on teaching and learning. The researchers examined the effect of weekly web-based quizzes on students' performance in a face-to-face psychology course at a public university in New England (Daniel & Broida, 2004). In one section of the course (*N* =44), the students did not have weekly quizzes, the students of the second section (*N* =42) took 15-minute pencil-and-paper weekly quizzes in class. The students in the third section (*N* =39) took the same 15-minute quizzes, in unproctored web-based format with automatically scored multiple-choice and short answer questions, during 24 hours before class. All three sections were taught by the same instructor, used the same lecture notes, and took the same four in-class exams at the end of each chapter. The mean of Exam 1 and Exam 2 scores for each group was analyzed at the middle of the semester ($M_{\textit{no quizzes}}$ = 48.95, $M_{\textit{web-based quizzes}}$

= 49.75, *M* *in class quizzes* = 59.45). A one-way-ANOVA showed a significant difference in the means across the groups (*F* (2, 122) = 46.69, *p* < .01, η^2 =.43). Bonferroni post hoc test demonstrated a significant difference in scores between the in-class quiz group and web-based quiz group (*t* (79) = 7.46, *p* < .001) and between the in-class quiz group and no- quiz group (*t* (84) = 9.38, *p* < .001). However, there was no significant difference between the no-quiz group and web-based quiz group (*t* (81) = .64, *p* > .05). To find possible explanations of why web-based quizzes did not have any influence on exam scores, the researchers distributed an anonymous survey in both quiz groups in which they asked the students to identify common cheating techniques during quizzes. Printing and sharing quiz questions, looking up the answers in the book during the quiz, using the glossary posted on the class website during the quiz, and taking the quiz in groups were listed by the web-based quiz group students (Daniel & Broida, 2004).

Based on the survey results, Daniel & Broida (2004) added 100 additional questions from the test-bank provided by the publisher to the online quizzes, activated randomization of questions feature, which randomly selected 10 questions out of 100 for each quiz, decreased the allocated time from 15 to 7 minutes, and remove the glossary from the site. The revised web-based quizzes were administered during the second half of the semester. A one-way ANOVA analysis of the means of Exam 3 and Exam 4 scores (*M* *no quizzes* = 46.45, *M* *web-based quizzes* = 60.90, *M* *in class quizzes* = 61.60) indicated a significant effect (*F* (2, 122) = 81.70, *p* < .01, η^2 =.57). Daniel and Broida applied Bonferroni post hoc test and found a significant difference in scores between the no-quiz group and web-based quiz group (*t* (81) = 11.47, *p* < .001) and between the in-class quiz group and no- quiz group (*t* (84) = 10.45, *p* < .001), and no significant difference between the in-class quiz group and web-based quiz group (*t* (79) = .54, *p* > .05). The researchers concluded that, with the use of randomization of quiz questions and reduced time, student performance on web-based quizzes can be equivalent to student performance during in-class quizzes, and better than student performance without weekly quizzes. Therefore, web-based quizzes, similar to in-class quizzes, have the potential to improve students' learning (Daniel & Broida, 2004). These results may be questioned because the researchers did not discuss whether the revised web-based quizzes were equivalent to in-class quizzes and did not specify whether the academic abilities across the groups were similar.

Daniel and Broida (2004) also noted that administering weekly quizzes outside of class brings additional advantages for teaching: instead of quizzing, valuable in-class time can be used for covering complicated concepts and problem-solving activities. Additionally, with automatic grading instructors do not need to spend their time on grading and recording web-quizzes' scores (Daniel & Broida, 2004). The researchers noted that further development of LMSs' testing features, including security mechanisms, can make administering of convenient and efficient web-based assessment more popular among faculty (Daniel & Broida, 2004). The findings of Lonn and Teasley's (2009) study described below reflected growth of the LMS and web-testing use over the years.

Lonn and Teasley (2009), who conducted their investigation in the spring of 2006 and spring of 2007, four years after Hanson and Robson (2004) and Daniel and Broida's (2004) studies, compared faculty and students' perceived benefits of the LMS Sakai at a large American Midwestern university. Sakai, an open-source platform released in 2004, incorporated latest technological innovations of that time (Dahlstrom et al., 2014). To measure the participants' perception, the researchers administered a survey with a 5-point Likert scale from strongly disagree to strongly agree (Lonn & Teasley, 2009). The study's results showed that 81% of 1,357 participating instructors in 2006 and 85% of 1,481 participating instructors in 2007 reported that they used the LMS in their classes. The researchers found that 47% of the instructors and 53% of 1,428 students perceived Sakai's online quizzes and exams as a valuable teaching and learning tool. According to the Chi-square tests, there was no significant difference in faculty and students' perceptions of the effectiveness of web-based exams ($\mathcal{X}^2$(2, 4888) =3.888, $p > .05$). However, students valued online sample and practice exams more than faculty ($\mathcal{X}^2$(2, 4577) = 37.507, $p < .001$) (Lonn & Teasley, 2009).

The wide-spread utilization of mobile phones and development of testing applications for mobile devices allowed for taking computerized tests on cellphones. In 2006, Romero, Ventura, and De Bra (2009) investigated whether students' performance on tests administered on cellphones is different from the students' performance on tests administered on PC computers. The researchers selected 30 computer science engineering students of the same age at Cordoba University in Spain (Romero et al., 2009). A simple multiple choice test with randomized items was created; the students had 1 minute to answer each question. The participants took the exams sequentially with a 30 minutes break in between: first the PC test and then the test on mobile phones. The mean of the time spent on each test, the number of skipped, incorrect, and correct answers were compared. The researchers applied a Tukey's Honestly Significant Difference and Ryan-Einot-Gabriel-Welsch tests and found that the time between PC and mobile exams differed significantly, although the difference was about 4.7 seconds ($M_{\textit{PC time}}$ =764.4 sec, $M_{\textit{mobile time}}$ =769.1 sec, p <.05; $M_{\textit{PC correct items}}$ =19.8, $M_{\textit{mobile correct items}}$ =18.9, $M_{\textit{PC incorrect items}}$ =6.2, $M_{\textit{mobile incorrect items}}$ =6.8, $M_{\textit{PC skipped items}}$ =4.1, $M_{\textit{mobile skipped items}}$ =4.9) (Romero et al., 2009). On the one hand a difference of 4.7 seconds may not look meaningful. On the other hand, Diedenhofen and Musch (2016) found that 3 seconds was enough to copy and paste the phrase from a test question into the search engine on the Internet. The results of a satisfaction survey, which was administered to the students after they took the tests, showed that students preferred to take the tests on PC computers with the bigger screen and better interface (Romero et al., 2009). However, they noticed, that mobile phones might be very useful for practice tests or self-assessment when PC are not available or not comfortable to use, for example, while waiting for public transportation. The researchers concluded that mobile phones can be used for simple multiple choice tests as a supplement in e-learning and suggested to study mobile testing with students of different majors and their teachers in future studies (Romero et al., 2009).

With further dissemination of web-based assessment in education, it was noted that computer testing may better accommodate students with special needs, including students with learning disabilities (Lee, Osborne, & Carpenter, 2010; Zenisky & Sireci, 2007). In 2008, Lee et al. (2010) conducted a mixed-methods study to investigate the benefits of computerized testing and influence of allocated time on performance of 31 university students with attention deficit hyperactivity disorder (ADHD). The participants were randomly assigned to one of four treatments: one group took computerized multiple choice exam with regular time, the time allocated for students without learning disabilities; the second group took the same computerized exam with extended time; the third group completed pencil-and-paper version of the same exam with regular time using scantrons, and the last one was assessed by the same pencil-and-paper exam but with extended time. The students with ADHD took all web-based exams administered through LMS Sakai at the university computer lab, which was proctored by an examiner. One minute per question was allocated in all exams with regular time, and 1.5 minutes per question was given in extended time exams. The web-exam items were displayed one at a time and were not randomized. Right after the exams, the students completed a written survey and participated in a follow-up face-to-face interview (Lee et al., 2010).

Lee et al. (2010) conducted a 2 x 2 factorial ANOVA and found that students performed significantly better on the web-based exams (*M pencil-and-paper* = 9.08, *M web-based =10.06*, *F* (1, 29) = 8.937, *p* = .014, *Cohen's d* =.3389). There was no significant difference between the scores with respect to allocated time across both types of exams (*M regular time* = 9.69, *M extended time = 9.61)*; however, the extended time had an effect close to significant during computerized exams (*F* (1, 29) = 4.102, *p*= .066). The analysis of the survey and interviews identified several advantages provided by computerized testing: students could read faster from the screen, did not need to worry about mis-bubbling the scantrons, it was easier and faster for them to select the answer instead of writing it, and one exam item at a time allowed for better focus. However, the participants in all groups were concerned with the pressure associated with a time limit. The researchers noted that the small sample size reduced generalizability of their findings, but recommended using web-based exams with sufficient allocated time as a viable accommodation for ADHD students (Lee et al., 2010).

In their case study conducted in Malaysia, Dahalan and Hussain (2011) investigated how web-based assessment features available in Moodle can be used to improve teaching and learning. Moodle is an open-source LMS, the first version of which was launched in 2002; Moodle 2, used in Dahalan and Hussain (2011) study, was released in 2010 (Moodle, 2012). Dahalan and Hussain (2011) utilized triangulation techniques involving documented analysis, observations, and interviews with three instructors and six students. All three instructors were trained how to build and use Moodle quizzes and exams in the classroom. The teachers created several short formative assessments with automatically scored multiple-choice, matching, and short answers questions and a practice final exam with 40 multiple-choice randomized questions to which they provided hints and general feedback. The Moodle adaptive mode option, which did not allow a student to start the next quiz question until the previous one was mastered, gave students the opportunity to learn in accordance with their abilities and pace. The instructors considered the formative web-based assessments and the practice final exam with randomized questions as the best way to identify the needs of their students. The faculty also mentioned that the use of web-based exams saved paper and reduced work-load associated with grading. The students valued immediate marking and provided feedback (Dahalan & Hussain, 2011).

In 2012, García-Cabrera, Ortega-Tudela, Balsas-Almagro, Ruano-Ruano, Peña-Hita, and Cuevas-Martínez (2012) examined the use of assessment tools and different types of exam questions available in ILIAS (García-Cabrera et al., 2012), an open-source LMS released in 2004 (Itmazi, Megías, Paderewski, & Vela, 2005). The researchers surveyed 250 professors of Jaen University in Spain, about 100 of whom responded to the survey questions. The survey results showed that 94% of the respondents used the ILIAS assessment tools for online homework and 64% of the participants administered web-based exams. About 86% of the instructor who used web-based exams utilized multiple-choice questions with a single answer, 56% incorporated multiple-choice questions with multiple answers, 43% used short answer questions, numeric questions, essay questions, and file-upload questions, and 21% included matching and image-map questions. The professors mentioned that they would like to add to the existing web-based testing tools an automatically scored item, which, similar to constructed response questions, allows for assessing problem-solving processes, not just the final result. The incorporation of interactive and multimedia-based exam questions was other desirable additional feature. The researchers concluded that further improvement of test security and the development of new types of web-exam exam questions that can help professors to evaluate their students more adequately may increase the effective use of web-based exams (García-Cabrera et al., 2012).

In 2012, Slepkov and Shielf (2014) suggested using a modified version of multiple-choice (MC) questions, an integrated testlet (IT), which allows for assessing some problem solving processes that cannot be measured by traditional (MC) items. The researchers converted several constructed response (CR) exam questions into multiple-staged MC items, which they called integrated testlets, and randomly assigned 155 Canadian introductory physics students to two sets of complementary midterm and final exams, in which each complementary exam had the same number of IT and CR items covered the same course material, but the question format was switched for each covered topic (Slepkov & Shielf, 2014). The CR questions on the midterm exam were graded independently by the researchers; two paid undergraduate students independently graded the final exams using the rubric (Slepkov & Shielf, 2014). The mean score on both versions of the midterm exam was 52%, and the mean scores on the two versions of the final exams were 51% and 52%. A correlation matrix that described the correlation between each exam question to every other question on the exam showed that IT and CR formats assessed the concept-equivalent questions similarly. Thus, although IT items are not entirely equivalent to CR questions with respect to measured knowledge, they assess procedural problem solving more similar to CR items than traditional MC questions (Slepkov & Shielf, 2014). Additionally, the researchers calculated that the cost of administering a CR exam is 20 times higher than the corresponding IT exam. Slepkov and Shielf (2014) recommended substituting CR items with IT questions for formal assessment in classes with large enrollment.

The next generation of computing technologies introduced automatic assessment of constructed-response questions, which has the potential to make more cost-efficient automatically-scored web-based testing entirely equivalent to pencil-and-paper exams with respect to adequate measuring of larger spectrum of students' knowledge (Smarter

Balanced Assessment Consortium, 2016). In 2013-2014, Smarter Balanced Assessment Consortium (2014) conducted a pilot test and field test studies of the previously mentioned Smarter Balanced Assessment which included automatically scored English and mathematics constructed response items. Each randomly selected constructed response question was independently scored by two human graders. Automated scoring systems provided by several vendors performed an additional scoring of the items graded by the humans. Quadratic weighted kappa (QWK) (Cohen, 1968) was used to measure human-human and automatic system-human scores agreement, which was considered satisfactory if QWK was at least .70 (Cohen, 1968; Smarter Balanced Assessment Consortium, 2016). An ANOVA of the mean QWK showed a significant difference between human-human scores agreement and automatic system-human scores agreement in each subject area, but in different direction ($p < .001$) (Smarter Balanced Assessment Consortium, 2014). For English constructed response questions, the best automatic scoring system had a mean QWK of .73 while human scoring had a mean QWK of .70. For mathematics, the best automatic scoring system had a mean QWK of .81 and human scoring had a mean QWK of .85. The lower mean QWK in automatic system-human scores agreement for math items could be explained by possible multiple representations of the correct responses nor recognized by the software (Smarter Balanced Assessment Consortium, 2014). The automatic scoring of constructed response items of web-based tests is improving (Smarter Balanced Assessment Consortium, 2014; Smarter Balanced Assessment Consortium, 2016) and may become widespread in several decades (Liu, Rios, Heilman, Gerard, & 2016; Smarter Balanced Assessment Consortium, 2016).

Interactive simulations, which allows for the assessment of skills not easily measurable during traditional pencil-and-paper exams, can be incorporated in web-based testing and bring new opportunities to engage students and improve learning (Redecker & Johannessen, 2013; Timmis et al., 2015). Bayes (2014) incorporated simulations into a formative online quiz constructed in Moodle to examine 49 Spanish high school students' understanding of the relative frequency distributions with different sample size in an introductory statistics course. The researcher created simulations using the dynamic open-source software Geogebra, administered the web-based quiz, and conducted follow-up interviews (Bayes, 2014). During the web-based quiz, the students were working on several tasks using the animated simulations to randomly generate data, analyze variations in the sample size, and compare different distributions and tendencies. The students reported that the visual aspect of the quiz simulations helped them to understand density of a distribution better, see how the sampling is done by the software, and compare different distributions. Bayes (2014) concluded that web-based testing simulations with clearly described successive tasks can improve students' understanding of the subject matter and actively engage them in reasoning about sampling distributions and modeling real data.

The advantages of computing, information, and communication technologies introduced during the last decade have been increasing popularity of web-based testing in higher education. In 2013-2014, in slightly over 10 years after their pilot investigation, EDUCAUSA conducted a study with 17,000 faculty and 75,000 students from US institutions to investigate the diffusion of LMS in higher education and explore the learning management systems' user experiences (Dahlstrom et al., 2014). The study's results showed that 99% of colleges and universities had at least one LMS which possessed a convenient and efficient way of delivering web-based exams. Over 85% of faculty responded that they used at least one available LMS for enhancing their teaching, including administering web-based assessment. The instructors valued flexibility in creating online tests, customized feedback, immediate automatic recording of exam scores in the LMS gradebook, and availability of test items' analysis provided by the system. About 83% of student-participants recognized the convenience of online tests and importance of immediate tests' feedback for their learning (Dahlstrom et al., 2014). Numerous higher education institutions have been renewing academic assessment using the latest innovations of computing, information, and communication technologies to evaluate, measure, and record academic learning (New Media Consortium, 2017). The community college where the study was conducted is one of them.

The Technological Advantages Used in the Study's Web-based Exams

Math faculty of the college where the study took place realized that web-based testing have many advantages and decided to implement web-based exams into the introductory statistics course curriculum in their web-assisted face-to-face, hybrid, and online classes. All web-based exams developed by the statistics instructors were administered through the college LMS Moodle 3.0 released in 2015. This version of Moodle incorporates many technological advantages of web-based testing described above. Moodle 3 has different types of automatically scored multiple-choice questions, including items similar to integrated testlets; automatically scored short answer questions and numeric questions with variable entries and units; and automatically scored interactive matching and drag-and-drop questions and image-map questions (Moodle, 2016a). Additionally, the testing features of this LMS contain manually scored essay questions, constructed-response questions, and file-upload questions, including audio and video files (Moodle, 2016a). Moodle 3.0 allows for using Fathom, Geogebra, and other similar dynamic software for statistical simulations (Moodle, 2016b). Moodle quizzes can be administered in standard or adaptive mode, with immediate or deferred feedback (Moodle, 2016a), and can be taken on mobile devices (Moodle, 2016c). Variable test time limit for an individual student, accessibility validation tools, compatibility with screen readers and other adaptive technologies make Moodle quizzes accessible by students with special needs (Moodle, 2016d). Moodle 3 automatically records exam scores, performs exam item analysis, and generates reports of students' performance (Moodle, 2016e). However, not all of these advantages were suitable for the given investigation.

Although adaptive mode, multiple-attempt integrated testlets, interactive Geogebra and Fathom simulations, file-upload and constructed-response questions are used in the course curriculum, these testing features were not included in the web-based exams involved in the given study. The department required all students to take identical exams with fixed time, which is not possible with adaptive mode and integrated testlets' "answer-until-correct" features. Constructed-response questions were not included because their automated scoring is not available in Moodle 3.0, and manual scoring may increase grading bias. Simulations and other interactive items may require extra time, manual assessment of skills needed for performing simulations (Raymond & Usherwood, 2013), and are more plausible for formative assessments or projects focused rather on learning than on testing (Bayes, 2014; Raymond & Usherwood, 2013). Thus, all four web-based exams involved in the study had automatically graded multiple-choice, short answer and drop-down (matching) items only. The study's web-based exams utilized the restricted time, automatic submission of exams when the allocated time expires, deferred feedback, blocked backtracking, and one test item at a time features. The individual time adjustment option was used for students with extended test time accommodations. The test items' marks, exam scores, and time spent on each test were automatically recorded by the LMS. Because it can take more time to complete a test on mobile devices (Romero et al., 2009), the students were asked not to use their cellphones for web-based exams.

Challenges of Web-Based Testing Associated with Cheating

Innovations in computing, information, and communication technologies brought web-based testing into the classroom (Ladyshewsky, 2015; O'Reilly & Creagh, 2016; Varble, 2014). However, at the same time, technological advances generated new and aggravated already existing challenges in all form of assessment, but especially during unsupervised web-based exams when students can cheat with ease (Bain, 2015; Moten et al., 2013; O'Reilly & Creagh, 2016; Shute & Rahimi, 2017). In addition to old-fashion crib notes, cheat sheets, and copying answers from a classmate, there are specific ways of academic dishonesty inherent to digital cheating during unsupervised online exams (Ravasco, 2012; Rogers, 2006, Shute & Rahimi, 2017). Among them are searching the Internet (Corrigan-Gibbs et al., 2015; Simpson & Yu, 2012), utilizing cell-phones to email or text friends (Bain, 2015; Rogers, 2006; Moten et al., 2013; Tindell & Bohlander, 2012), using multiple accounts to get access to exam answers (Northcutt et al., 2016), or emailing online companies that can answer the question or take the entire tests for a student (Malesky, 2016). Thus, technological advances create favorable conditions for cheating.

Possible collusion during unproctored online exams is an ongoing concern of educators (Faurer, 2013; Pittman, 2015; Ravasco, 2012; Rogers, 2006; Shute & Rahimi, 2017; Srikanth & Asmatulu, 2014). Collusions can be categorized as unauthorized collaborations, helping each other in answering test questions; unauthorized coaching, responding to test questions with unauthorized help of a tutor; and paid-unpaid surrogate, when the entire test is performed by another individual (Trenholm, 2007). Through the use of social networks, students can share files with solutions, discuss answers using personal chat rooms, and participate in group cheating when the students assign exam questions to each other and share the answers (Ravasco, 2012; de Sande, 2015). While widely-used technology-based services similar to Turnitin can be utilized to identify some aspects of academic dishonesty during writing-oriented assessment, such identifiers for fact-based online exams do not exist (Faurer, 2013; Trenholm, 2007). For these reasons, many instructors are not convinced that the level of integrity of unsupervised web-based exams can be comparable to the level of integrity of proctored tests (Fask et al., 2015; Rogers, 2006; Trenholm, 2007) and may refuse administering unsupervised web-based exams or teach fact-based online courses (Bandyopadhyay & Barnes, 2014; Rogers, 2006). Thus, difficulties in controlling any form of unauthorized help during unproctored exams may negatively influence adoption of web-based testing and e-learning. Several researchers studied academic dishonesty on online exams to determine whether cheating takes place, in what forms, and to what extent. Their studies focused on empirical evidence of cheating, unauthorized collaboration, and the use of the Internet and other technological advances for cheating purposes on online tests.

Studies Related to Empirical Evidence of Cheating on Web-Based Exams

In 2004-2005, Harmon and Lambrinos (2008) studied likelihood of cheating during unproctored exams in their quantitative study with online microeconomics students at the University of Connecticut. This was one of the first quantitative studies which utilized empirical methods to measure likelihood of cheating (Beck, 2014; Brothen & Peterson, 2012; Harmon & Lambrinos, 2008). One online section of the course (*N* =24) was offered in summer 2004; the second online section of the course (*N* =38) was offered in summer 2005. Both sections were taught by the same instructor (O. Harmon, personal communication, March 5, 2015), who utilized the same materials and web-based exams delivered through WebCT LMS (Harmon & Lambrinos, 2008). However, the first section had an unproctored web-based final exam while the second one had the same web-based exam, but proctored. This cumulative 90-minute final exam had 30 multiple-choice items randomly selected from the pool of 100 questions. The proctored group of students could use notes, books, and calculators, but not cell-phones; a verbal cheating warning was given to this group. The three chapter exams, the final exam scores, the students' GPA, age, major, and college grade level were used for the OLS analysis (Harmon & Lambrinos, 2008).

The series of independent t-tests showed that there were no significant differences in students' GPA and final exam scores across the groups (M_{GPAup} = 2.86, M_{GPAp} = 3.00, t = -.99, p > 0.05, Mup final exam = 73.23, Mp final exam = 77.15, t = -1.32, p > 0.05) (Harmon and Lambrinos, 2008). Although the proctored group performed better on all three chapter exams than unproctored group, the difference was significant only for Exam 3 (Mup exam 3 = 68.75, Mp exam 3 = 78.09, t = -2.89, p < 0.05). Harmon and Lambrinos (2008) inferred that the students' academic abilities were similar. The comparison of R^2 -statistics of OLS models for both groups, which the researchers used to measure likelihood of cheating, suggested that cheating took place during unproctored final exam because the human capital variables explained much smaller variation in exam score during unproctored exam than on proctored one (F_{up} =.02, R^2 up =.497, p > .05, F_p =35.60, R^2 p =.0008, p < .01). The researchers concluded that, although the sample size was small, they found some empirically supported evidence of cheating during unproctored exams (Harmon & Lambrinos, 2008).

Carstairs and Myors (2009) compared undergraduate industrial and organizational psychology students' performance on proctored and unproctored high-stake cognitive tests. Group 1 (N =159) took a proctored pencil-and-paper 55 multiple-choice item exam. In the following year, Group 2 (N =143) completed the same exam in parts: the first 20 questions were given as a take-home test, the middle 20 questions were administered as unproctored web-based exam through WebCT, and the remaining 15 items were incorporated into the proctored pencil-and-paper final exam. The web-based exam was open for 10 days, had infinite number of attempts, and was untimed. Both groups were similar with respect to age, GPA, and gender. The researchers planned to administer the first part of the test as a web-based exam too, but were not able to build it in the LMS on time. The exams' scores in both groups were analyzed (Carstairs & Myors, 2009).

A 2 x 3 mixed ANOVA, with the between-subject factor Group and the within-subject factor Test, showed significant main effect of Group (*F*(1, 300) = 114.84, p <.001), Test (*F*(1, 600) = 227.56, p <.001), and significant Group x Test interaction (*F*(1, 600) = 113.56, p <.001) (Carstairs & Myors, 2009). Thus, performance during unproctored exams was significantly better than during proctored exams (M_{G1T1} =12.31, M_{G2T1} =17.35, *d* =2; M_{G1T2} =14.25, M_{G2T2} =17.43, *d* =1.21), but the performance was not significantly different between unproctored exams (M_{G2T1} =17.35, M_{G2T2} =17.43). Both groups performed similarly on the third proctored test (M_{G1T3} =12.31, M_{G2T3} =12.15, *d* =-.05), which demonstrated their similar abilities. The researchers concluded that the effect appeared due to unproctoring, but not the format of the exam: the findings provided empirical support that cheating during unproctored web-based exams took place. The authors recommended avoidance of unproctored high-stakes cognitive tests without finding effective ways to prevent cheating (Carstairs & Myors, 2009).

In 2011, Brothen and Peterson (2012) conducted a natural experiment at a large Midwestern university when the WebVista LMS crashed after the first 30 minutes of the proctored web-based final exam in Theories of Learning course. The students (*N* =25), who could not finish the exam because of the crash, were told to take the same 90-minute web-based exam with 82 randomized multiple-choice items at home, but without books, notes, and other help. These students constituted the quasi-experimental group, while other 173 students, who took the web-exam in the proctoring center scheduled on the different days, represented the control group. The researchers applied the independent t-test and found no statistically significant difference in exam scores between the proctored and unproctored groups (M_{up} = 53.76, M_{p} = 47.86, t =3, p < .01). Moreover, the unproctored group spent on average 25 minutes more than proctored one, and this difference was significant (M_{up} = 76 min, M_{p} = 51 min, t = 5.97, p < .001). Longer testing time could indicate that the students were looking up the needed information in the book or other resources. To test whether the control and experimental groups differed in academic performance before they took the final, the authors compare the students' average scores on other course assignments and found no significant difference (M_{up} = 105.16, M_{p} = 102.20, t =.77, p =.44). Brothen and Peterson (2012) concluded that although better performance on the unproctored final exam and longer test taking time could suggest that some cheating took place during unproctored exams, other factors, which were not possible to control in the natural experiment, could have influenced the students' performance. To prevent possible cheating, the researchers recommended the implementation of realistic time limits (Brothen & Peterson, 2012).

Olivero (2013) investigated the relationship between cheating and testing time by comparing test times and scores on unproctored online exams at a university in the Northwest. The university adopted the honor code and implemented anticheating policies. There were four groups of undergraduate students in a foundational survey course: two classes (*N1* =59, *N2* =79) had no time limit on all three exams and two classes (*N3* =56, *N4* =54) had time limit on the exams. Most of the participants were juniors (87%). The same 40-item multiple-choice tests were administered to all groups through Blackboard. Each test item was presented one at a time, the backtracking was blocked, and the Respondus Lockdown browser prevented the use of other sites during the exam. However, the questions were not randomized and the students could see whether they responded to the problem correctly, although the correct answers were not provided. The time spent by students on each of the three tests and the exams' scores were analyzed (Olivero, 2013).

Olivero (2013) utilized a one-way ANOVA and found a significant difference between the unlimited time and timed groups during all tests ($M_{T1untimed}$ =60 min, $M_{T1timed}$ =32.5 min, F =28.74, p <.000; $M_{T2untimed}$ =67.5 min, $M_{T2timed}$ =26.5 min, F =57.40, p <.000; $M_{T3untimed}$ =67.5 min, $M_{T3timed}$ =27.5 min, F =25.88, p <.000). The mean score analysis demonstrated that the untimed group performed statistically better than the timed one ($M_{T1untimed}$ =326.5, $M_{T1timed}$ =302.5, F =4.52, p <.000; $M_{T2untimed}$ =324, $M_{T2timed}$ =292.5, F =8.93, p <.000; $M_{T3untimed}$ =342.5, $M_{T3timed}$ =301.5, F =24.379, p <.000). Olivero explored what percent of students spent time on exams below and above alpha of .005 in comparison with the first timed group. About 75% of the participants in untimed groups were in this range. The author inferred that the untimed groups cheated by using more time to consult the textbook or notes, which could result in better exam scores. The students who had time below alpha .005 most likely were getting answers from their friends. The researcher concluded that the findings suggested that, if an opportunity is present, students are cheating during online unproctored exams in spite of the honor code and cheating policies stated in the syllabus (Olivero, 2013).

Beck (2014) investigated evidence of cheating by comparing students' performance during proctored (*N* =81) and unproctored (*N* =19) exams in the introductory economics university course. There was no significant difference with respect to GPA and other available human capital variables across the sections (p >.05). The proctored group took the pencil-and-paper exams on campus, while online students took a web-based version of the same exams administered through D2L LMS (V. Beck, personal communication, May 17, 2015). The questions on web-based exam were randomized, only one question on the screen was presented, the backtracking was blocked, and the answers and feedback were not provided until all students took the test. The cheating warning was given before each exam in both groups (Beck, 2014).

The t-test demonstrated no significant difference in scores on proctored and unproctored exams (M_{pr} =40.21, M_{up} =40.63, t =.347, p >.05) (Beck, 2014). Several OLS regressions indicated that the R^2 statistics explained less variation in the exam scores in proctored group than in unproctored group ($R^2\ pr$ =0.197, $R^2\ upr$ =0.331). According to Harmon and Lambrinos (2008), this fact could indicate that cheating occurred in the proctoring section, which seemed unlikely. Beck (2014) explained that the difference in her findings and Harmon and Lambrinos' (2008) results could take place due to the fact that her investigation utilized more cheating prevention technics than Harmon and Lambrinos' (2008) study, which incorporated only randomization of test items and warning statements. Beck (2014) noted that, although her study did not indicate evidence of cheating on unproctored exams, academic dishonesty may take place in all form of assessments and should be controlled by the best available security mechanisms.

Sivula and Robson (2015) compared students' performance during proctored and unproctored final exams in the face-to-face (*N* =20) and online (*n* =21) sections of the graduate research methods course. The proctored group had two hours to complete a comprehensive final exam, which included 22 short answer essay questions with sub-parts. The unproctored group took the web-based version of the same final exam administered through Blackboard. The online exam was open for four days, had multiple attempts, and unlimited time and resources. However, the students were warned to complete the test individually without any human assistance. The instructor graded both exams using the same grading rubric. The t-test indicated a significant difference in students' performance on proctored and unproctored exams (*Mup* =90.18, *Mp* =80.10, *t* =3.40, *p* =.001, *Cohen's d* =1.084); the corresponding CI did not include 0 (95% CI [4.04, 16.11]). The unproctored group performed 34% better than proctored one, indicating that cheating took place during the online exam (Sivula & Robson, 2015). It can be said that Silvia & Robinson (2015) study's results might have limited generalizability due to the small sample size. However, Brallier and Palm's (2015) investigation with bigger sample size described below also demonstrated evidence of cheating.

Brallier and Palm (2015) compared students' performance on proctored and unproctored exams in undergraduate introductory sociology course at a southeastern university. The data for the study were collected during four consecutive fall semesters. Each semester, the same instructor taught two sections of the same course. During the first two semesters both sections took the unproctored web-based multiple-choice final exam (*N* =130), while during the following two semesters both sections took the proctored pencil-and-paper version of the same exam (*N* =116). At the beginning of each term the students took a 50 multiple-choice web-based exam to identify the baseline knowledge of sociology concepts. The same materials, assignments, and tests were used in all sections. A 50-minute web-based open-book and open-notes final exam with randomly assigned questions was administered through WebCT/Blackboard; the students had 24 hours to complete the exam (Brallier & Palm, 2015).

Brallier and Palm applied ANOVA and found that the test format had a significant effect on exam scores: the unproctored group scored significantly higher than proctored one (*Mup* =74.66, *SD* =10.87, *Mp* =68.65, *SD* =12.12, *F* (1,242) =17.41, *p* <.001, η^2 =.07) (Brallier & Palm, 2015). The independent t-test demonstrated no significant difference between the proctored and unproctored groups with respect to GPA and pretest scores (*MGAup* =*MGPAp* =3.04, *Mpreup* =52.72, *Mprep* =49.91, *t* =1.73, *p* =.09).The researchers concluded that their findings suggested that cheating took place during unproctored exams: on average, the unproctored group scored 6% higher (Brallier & Palm, 2015).

Fask et al. (2015) examined evidence of cheating during unproctored exams in undergraduate introductory statistics course at a small private university in the Northeast. The convenient sample of 52 students took the final exam in two parts: the unproctored web-based online (Fask et al., 2015) and proctored pencil-and-paper in-class (F. Englander, personal communication, November 28, 2015). The questions for both exams were taken from the same test bank provided by the publishers and were assumed to be equivalent. Both 2-hour exams, which were open-book, tested students' ability to solve numerical statistical problems and did not include multiple-choice and essay questions. On both exams, the students were asked to solve five problems and had about 24 minutes per problem (F. Englander, personal communication, November 28, 2015). To obtain empirical evidence of cheating, the researchers intentionally did not use any cheating prevention mechanisms: the web-based exam questions were not randomized; the backtracking and other options used by Beck (2014) were not activated (F. Englander, personal communication, November 28, 2015). The web-based exam was administered first through Blackboard and had three-day window; the students took the equivalent proctored exam right after that (Fask et al., 2015). Fask et al. (2015) noted that because the unproctored test was administered first, the possible testing effect could create a measurement bias against evidence of cheating on the unproctored exam, but not the other way around.

To detect cheating, Fask et al. (2015) used the latent variable approach. The descriptive statistics showed that the students performed better on the unproctored exam (*Mpr* =65.14, *Mup* =72.96). The researchers used the Blackboard score distribution for each test item to grade the pencil-and-paper exam the same way (F. Englander, personal communication, October 21, 2016). The model that depicted the cheating process was tested in software SAS using a maximum likelihood function and the covariate matrix as an input (Fask et al., 2015). The researchers found evidence of academic dishonesty: the path diagram connected the latent variable *cheating* and observed variable *score on unproctored exam* with the regression coefficient of 1, indicating that cheating had a direct effect on exam score during unproctored test. Fask et al. (2015) recommended investigating the effect of order in which the proctored and unproctored exams are administered in future research and incorporating effective combinations of cheating prevention methods.

Arnold (2016) studied empirical evidence of cheating on unproctored web-based exams with a cohort of 461freshmen who were pursuing the bachelor degree in economics at Erasmus University of Rotterdam in the Netherlands. The students in this cohort took four courses, Microeconomics, Statistics, Accounting I, and Accounting II, with unproctored formative exams and two courses, Mathematics I and Mathematics II, with proctored formative exams. The unproctored exams consisted of randomized multiple-choice questions and had restricted time. All courses had proctored summative exams at the end of the term. The average scores on unproctored exams ranged between 7.75 and 8.34 while the scores on the proctored exams were between 4.32 and 7.49.

To detect cheating, Arnold (2016) applied Harmon and Lambrinos' (2008) OLS model and Jacob and Levitt's (2003) algorithm for identifying unexpected fluctuations in test scores. The R^2 for the unproctored tests ranged from .0096 to .176 while for the proctored tests this statistic was between .150 and .254. According to Harmon and Lambrinos' (2008) model, the lower R^2 could indicate evidence of cheating during unproctored exams. Jacob and Levitt's (2003) algorithm provides a formula to calculate Jacob and Levitt's scores (Arnold, 2016). Positive correlation coefficients between Jacob and Levitt's scores across the groups suggest high likelihood of cheating. The correlation coefficients in unproctored group ranged between 0.35 and 0.46 (p <.01), while the correlation coefficients in proctored group were between -.02 and .11 (*p* <.05), indicating that cheating occurred during unproctored exams. Additionally, direct observations of the students scored showed that seven students had extremely high scores on the unproctored tests (around 9.5-10 out of 10 possible) and extremely low scores (around 1-2.3) on the summative proctored exam. Arnold (2016) concluded that the results of the study suggested that cheating during unproctored exams took place.

The studies described above found some evidence of cheating during web-based exams. However, they did not investigate how students can cheat. The section that follows describes research with emphasis on one form of cheating during unproctored exams, unpermitted collaboration.

Studies Related to Unauthorized Collaboration on Web-Based Exams

A friend's assistance during online exams was studied by Gustafson (2002) with 170 Macroeconomics Principles university students. The researcher examined how an "open friend" policy influenced students' behaviors and exam scores during ten unproctored web-based exams administered outside of the class. The students were allowed to take the tests any time and use any external resources including books, classmates, and other friends. Each exam had limited time and randomly selected questions from a test bank of 150-300 questions. It was assumed that frequent testing, limited time, and randomization of exam items could not allow for extensive collaboration during the online tests. At the end of each online exam the students received extra credit for answering additional questions about whether they were taking the test with a friend or not (Gustafson, 2002).

Ninety-five students (56%) reported that they took all 10 tests individually, 17 students used a friend for all tests, and the remaining 58 students used their friends' assistance periodically (Gustafson, 2002). Forty-one participants utilized the same friend's assistant, 22 students were helped by two different friends, 11 students got help from three different students, and two students were helped by four different friends. Six students had assistants who were not their classmates. All students who got help from their classmates took the tests in pairs: when the first student was taking the test, the second student was helping him. When the second student was taking the test, the roles were switched. The seemingly unrelated regression analysis showed that friend's assistance had a weak positive, but statistically insignificant effect on a student's test score (p <.01). By the end of the semester, the number of students who took the tests with friends decreased by 29%. The researcher concluded that friend's assistance during online exams is a reality, but can be reduced by implementing thoughtful curriculum design, utilization of restricted time and randomization (Gustafson, 2002).

Another study on friend-based cheating was conducted by Chapman, Davis, Toy, and Wright (2004) who examined student behaviors and perceptions with regards to cheating on unsupervised web-based exams in their two-stage study at a midsized Western university. During the first exploratory stage, the researchers held two discussion groups, 20 students in each, to identify the major themes related to web-based exam cheating. The discussions in both groups showed that the participants did not believe that sharing information about test questions with friends who had not yet taken the exam is cheating. Many respondents said that helping friends to solve problems during proctored or unproctored exams is not unethical. Friend-based cheating became one of the major components of the survey distributed to 824 business students during the second stage of the study (Chapman et al., 2004).

The analysis of the survey responses demonstrated that about 88% of the participants believed that working together with a friend on an unsupervised web-based exam was cheating, but would help their friends during exams anyway. Only 58% thought it was cheating to share information about the exam even if professor clearly prohibited it. About 24% of the students self-reported that they have cheated at least once on an unsupervised web-based exam, 2% (about 16 students) reported that they worked together while taking an e-test, and 42% said that they would cheat during an online exam if they had the chance. The researchers noted that the actual number of unauthorized collaborations on exams could be much higher than was self-reported and suggested utilizing limited time with randomization of test items and postponing posting the exam answers to make collusion among students more difficult (Chapman et al., 2004). A similar concern about unpermitted collaboration was raised in Stuber-McEwen, Wiseley, and Hoggatt's (2009) study described below.

Stuber-McEwen et al. (2009) explored self-reported incidents of cheating of 87 face-to-face and 138 online students at a private mid-size Midwestern metropolitan Christian university. The survey results showed that about 45% of face-to-face and 10% of online students reported being involved in at least one form of academic misconduct. A 2 x 2 Chi Square test indicated a significant difference in overall cheating between face-to-face and online students (χ^2 =33.75, p <.0001). The online group of students was older than face-to-face group, which could explain the significant difference (Stuber-McEwen et al., 2009). Out of all students who self-reported cheating, 41% of face-to-face students and 43% of online students admitted cheating on exams (*p* =.0013). Collaboration, abetting and aiding, was the most commonly reported type of cheating in both groups with 50% among online and 69% among face-to-face students with self-reporting cheating (*p* =.0001). The researchers were surprised that abetting and aiding appeared to be the most frequent form of self-reported academic misconduct and suggested to find ways to reduce unpermitted collaboration in all course delivery modes (Stuber-McEwen et al., 2009).

The further development of LMSs allowed for analyses of not self-reported evidences of friend-based cheating during web-based exams. In 2010-2014, a team of researchers studied unauthorized collaboration during online tests in Signals and Systems, a mandatory course for electrical engineers at the Technical University of Madrid, Spain (de Sande, 2015). In 2010, the researchers created a 5000 item-bank in Moodle, which consisted of automatically-scored calculated, also called numerical, questions with randomly-assigned numerical parameters of the same test item. From this test bank, the LMS randomly generated 10 questions for an exam with 100 different variations for the numerical parameters of each test-item. The questions were at analyzing level of Bloom’s taxonomy and required inserting the answer in the provided space. The test was open-book and open-notes and could be accessed any time during two to four days period, but the students were required to take the exam individually (de Sande, 2015).

The first time the test was administered in 2010-2011 academic year; however, after two consecutive semesters of using the same test bank, the researchers noticed that many students solved the exam problems correctly in suspiciously small period of time. In 2011-2012 academic year, 23 out of 102 students (22.5%) solved all 10 problems in less than 12 minutes and obtained over 9 points out of possible 10. The average score for that semester was 7.4, which led to an assumption that about 1/5 of the class was getting unfair help. The number of students who did suspiciously well was larger in the following year: 40 out of 168 students (23.8%) obtained scores of above nine in less than 12 minutes during 2012-2013 academic term. The average mark for that year was 6.8 points. The authors hypothesized that students shared the numerical test items with the solutions to them among each other. To test this hypothesis, the researchers added several new questions to the test-bank for the 2013-2014 academic year and programmed LMS such that when a student opened a test, Moodle randomly selected just one new item. The date, the time at which each test was opened, the number of minutes spent on each question, and whether the item was answered correctly or not were analyzed (de Sande, 2015).

Seventy-six students took the exam in 2013-2014 academic year. The analysis showed that the fastest students had a score around eight, which was lower than in two previous years (de Sande, 2015). The first 17 students in a row (22.4%) got the wrong answer to the new question, the correct response to which appeared in six hours after the first student opened the test. About 45% of the student solved the new question correctly right after the appearance of the correct answer. This pattern suggested that students who took the test first did not know about the new questions and used the solutions provided by previous years' students; however, in six hours the solutions to the new problems were found and disseminated among the rest of the interested students. The author concluded that although numerical questions at the higher cognitive levels, large test banks, and randomization of exam items can be helpful in reducing cheating on quantitative online exams, these techniques do not prevent from students' collaboration on creating solutions to the exam questions and combining them, for example, into a Google spreadsheet which can be instantaneously accessed by any individual interested in cheating. Further research on methods which provide secure and credible unproctored web-based assessment is needed (de Sande, 2015). In addition to unauthorized collaboration during online testing, researchers studied academic dishonesty that can occur through the use of technology.

Studies Related to Unpermitted Use of Technological Advances on Online Exams

Rogers (2006) conducted a survey-study to examine faculty perceptions about e-cheating on web-based tests administered through WebCT LMS at a southeastern US university. Twenty two out of 54 respondents (41%) did not use WebCT testing features: 36% of these 22 instructors listed cheating during exams as a major reason for not using web-based testing. The instructors who used web-based exams in their classes also expressed a concern about cheating and shared their experiences with it. About 53% of those who used web-based exams reported at least one detected occurrence of academic dishonesty. The most frequent cheating method during web-based exams observed by the instructors was looking at neighbor's computer (22%), sending emails (19%) and searching the Internet (13%). Other methods included instant messaging (9%), cell phone and text messaging (9%), using not allowed files or software (9%), and accessing unauthorized sites (3%) (Rogers, 2006). Thus, Rogers' (2006) findings indicated that students use technology to cheat on web-based exams.

Another study on self-reported evidence of cheating on web-based exams was conducted by McCabe et al. (2012) at Texas Tech University in 2010 with 1,043 student and 479 faculty respondents. About 60% of participated students reported that they have taken an online exam; 22% of participated faculty reported offering an online exam. About 19% of these students and 53 % of the instructors observed unpermitted collaboration during an online test; 27% of the students and 41% of the faculty noticed the use of books and notes; 14% of the students and 36% of the instructors observed receiving unauthorized help; and about 27% of the students and 40% of the faculty reported the incidents of unpermitted search of the Internet (McCabe et al., 2012).

Self-reported e-cheating during web-based exams was also studied by Khan and Balasubramanian (2012) with 224 students from different universities in the United Arab Emirates. About 78% of the participants reported being involved in some form of cheating through the use of technology. Over 35% of the respondents searched the Internet, 19% used programmable calculators for cheating purposes, 10% utilized mobile phones, 10% used i-pods, 9% had memory sticks with unpermitted information, and 9% purchased needed resources or solutions online. Khan and Balasubramanian concluded that e-cheating is a common behavior in higher education and should be neutralized.

Tindell and Bohlander (2012) specifically studied the use of cell phones and text messaging during lectures by surveying 269 college students at a small private university in Pennsylvania. Two hundred sixty six out of 269 (99%) said that they brought a cell-phone to school every day. Over 33% of the respondents observed another student using a cell phone for texting during exams, and about 10% of the respondents self-reported that they text their classmates during an exam. About 99% of all respondents believed that they could text undetected with the class size of over 100 students. Tindell and Bohlander concluded that the use of cellphones for cheating purposes is an issue which cannot be ignored.

Unlike Rogers (2006), McCabe et al. (2012), Khan and Balasubramanian (2012), and Tindell and Bohlander (2012), who utilized potentially unreliable self-reports, Simpson and Yu (2012) analyzed e-cheating via electronic records provided by the LMS. Thirty-six introductory psychology lab undergraduate students at a small southeastern private college with honor code policy participated in the study (Simpson & Yu, 2012). Several 10-minute web-based quizzes with four multiple choice and one short answer questions were created and administered online through Blackboard, which preserved the detailed log of all activities on any computer (Simpson & Yu, 2012). All quizzes, which contributed less than 25% to the overall course grade, were taken in the unsupervised computer lab. In Simpson and Yu's (2012) study, the students were not allowed to use books, notes, cell phones, or Internet.

The analysis of the activity logs showed that the posted online resources were accessed via the Internet during the quizzes six times (8.3%) (Simpson & Yu, 2012). A previously administered survey at the same institution about using the Internet during online exams identified 5% self-reported incidents of this cheating behavior. The researchers concluded that self-reported data have a tendency to underestimate actual occurrence of cheating. Simpson and Yu (2012) also noted that, in spite of an institutional honor code and academic integrity oath, cheating during unsupervised exams still occurs, even if the quizzes do not influence the overall course grade a lot. On the other hand, the investigation did not demonstrate too frequent use of the Internet either. Simpson and Yu recommended conducting other studies on academic cheating during unsupervised web-based exams that are not based on self-reports. One of such investigations was conducted by Corrigan-Gibbs et al. (2015) in 2014.

Corrigan-Gibbs et al. (2015) studied the use of the Internet during an unsupervised web-based exam with 404 undergraduate engineering students in India who took a MOOC "The Design and Analysis of Algorithm." The online final test for this course, the successful passing of which qualified for a certificate and a possible internship position at Microsoft Research India, had one free response and 15 multiple-choice questions. All questions were original and required critical thinking; thus, it was unlikely to find solutions to them on the Internet. To decrease unpermitted collaboration, 15 different versions of the same exam were created and randomly distributed to the students through the LMS; the questions and answers to the multiple-choice questions were also randomized. Although the exam could be completed in 1 hour, the test window was open for 2.5 hours. The students were not allowed to use books, notes, the Internet while taking the test and neither receive or give help to other people during the exam (Corrigan-Gibbs et al., 2015).

To detect cheating and check whether the students were searching the Internet during the exam, Corrigan-Gibbs et al. (2015) created a “honey pot” Google website where they placed all of the exam questions. The website had a button “Show answers,” but the actual answers were not provided. The researchers used the cookie and tracking iframe of the LMS to identify the students who visited the website and clicked on the button “Show answers” during the exam. Additionally, the authors isolated specific language and symbols in the free response answers and inserted them into the Internet search engine for detecting plagiarism. To identify copying among the students, the answers to the free response question were carefully and independently examined by the three of the researchers. The assigned scores agreed with each other on 95% of the responses; Cohen’s Kappa statistics of *k* =.83 showed almost perfect inter-grader agreement (Corrigan-Gibbs et al., 2015).

The analysis showed that 100 students (25%) cheated on the final exam (Corrigan-Gibbs et al., 2015). The researchers identified 23 IP addresses visited the “honey pot” website; however, five of them were not associated with the examinees’ addresses recorded in the LMS. Corrigan-Gibbs et al. (2015) suggested that five students could use their cell phones or other devises to search the Internet, but not the computers on which they were taking the test. Thus, 18%-23% out of identified 100 “cheaters” visited the “honey pot” web-site, about 84% plagiarized on the free response question, and 2% did both. Similarities among the students’ answers to the free response question were determined in 80% of the plagiarized incidents, 42% of responses were identical to Internet websites, and about 20 % were similar to online resources and other students' answers. The students plagiarized using Wikipedia, a tutorial on how to design a graph algorithm, and peer-reviewed articles. The same excerpt from Wikipedia was used by eight students; the same sentence from the same peer-review publication appeared in the answers of two students. None of the plagiarized answers was correct: the "cheaters" performed significantly worse than other students ($M_{cheaters}$ =30% correct, $M_{not\ cheaters}$ 40% correct, t = 4.18, p <.0001). Therefore, the students who cheated were academically weaker than the students who did not cheat. Corrigan-Gibbs et al. concluded the test questions with emphasis on critical thinking, different versions of the exam, and randomizations of the test items did not allow the students who cheated to get unfair better scores. Another copying cheating strategy used in MOOCs was examined in Northcutt's et al. (2016) investigation described below.

Northcutt et al. (2016) studied the use of multiple-accounts during online certificate exams administered in MOOCs. Multiple account users create one or more "harvester" MOOC accounts to access a test's answers, copy, and paste them into the same test taken through the major account (Northcutt et al., 2016). To detect multiple accounts' users, the researchers examined 189,092 accounts in 115 MOOCs offered through edX platform at Harvard and MIT in 2012-2015. The analysis based on a Bayesian criterion detection algorithm showed that about 1,237 certificates were earned through the use of multiple accounts by 657 students who utilized 674 harvester accounts. About 1.3 % of certificates were earned through multiple accounts detected in 69 courses

(60%). In some of these courses, up to 5% of the certificates were obtained through the multiple accounts cheating strategy. The biggest number of unfairly earned certificates (1.2%) was identified in Government and Social Sciences courses, the instructors of which did not employ any cheating prevention methods. The smallest number of unfairly earned certificates was detected in 18 STEM courses (0.1%) where the test questions were randomized and the answers to them were not accessible until the exam was due. The researchers concluded that the honors codes widely used in MOOCs is not enough to prevent cheating and called to find other ways to increase credibility and trustworthiness of MOOC certificates (Northcutt et al., 2016).

Malesky et al. (2016) conducted an investigation about a cheating company that completes mathematics, statistics, accounting and other online courses for college and university students. A graduate student, who played the role of a fake online Introductory Psychology student, contacted a company through a website. The company completed the entire course, including all quizzes, exams, discussions, and projects, and received an A for the student. The company's representative refused only to do the oral online live presentation at the end of the term, but prepared all needed materials for the student. All course papers were submitted through Turnitin, all answers to exams were carefully checked for copying; however, both professors, who were co-teaching the course, did not notice that the student was "cheating." Malesky et al. concluded that, although most students cannot afford paying $900-$2800 per course for cheating companies' services, educational community should be aware of such type of academic dishonesty and find ways to prevent it.

Cheating Prevention Techniques

The previously discussed studies have demonstrated that e-cheating is an issue. Educational and research communities called to find ways to reduce academic dishonesty and increase the credibility of web-based testing to a level comparable to proctored exams (Corrigan-Gibbs et al., 2015; de Sande, 2015; Faurer, 2013; Malesky et al., 2016; Northcutt et al., 2016; Pittman, 2015; Trenholm, 2007). In 2009, the WCET, which consists of 16 states, including California where the study took place, published a document about best practices in promoting academic integrity, which can be divided into two major categories: institutional cheating prevention strategies and security mechanisms available in LMSs. Student logins and passwords to access school web-sites, campus-wide policies on academic misconduct, honor code, cheating statement with clearly explained repercussions stated on a syllabus, warning statements posted on websites, and proctoring are some of the institutional academic integrity strategies. Randomization of exam questions, forced completion of the test, limited test time, displaying exam items one at a time, lock-down browser, higher order thinking exam questions, and deferred feedback were listed in the LMS security mechanisms category (WCET, 2009). Previous research with emphasis on institutional cheating prevention methods related to the given study is described below.

Studies Related to Institutional Cheating Reduction Techniques

King et al. (2009) studied 121 undergraduate online accounting students' perception of cheating with respect to the test-taking cheating policies stated on the syllabus. A Likert-type scale data analysis showed that after the cheating behaviors on unproctored tests were identified on the syllabus, students' perception that getting help from other individuals and utilizing online sources during exams is cheating increased from 50 % to 82% and from 21% to 78% respectively. King et al. concluded that clearly stated cheating policies have the potential to neutralize academic dishonesty.

Staats and Hupp (2012) examined whether self-reported cheating intentions of 325 psychology students in a Midwestern university were influenced by the syllabus cheating statements. The researchers randomly assigned the participants to three groups: the control group (*N* =89) without cheating statement on the syllabus, the group (*N* =129) with the university academic misconduct policy stated on the syllabus, and the academic integrity group (*N* =92), in which the syllabus stated that cheating makes education pointless. After each experimental group was exposed to the corresponding statement on the syllabus, they were asked to answer a series of Likert-type questions, one of which was whether the statement on the syllabus discouraged the participants from cheating. An ANOVA showed no significant difference in intention to cheat across the groups (*ps* >.05; F-statistics was not provided), indicating that the syllabus statement did not influence cheating intentions. Staats & Hupp concluded that cheating statements on a syllabus may clarify which behavior constitutes academic misconduct, but they alone do no prevent from cheating: a combination of numerous cheating prevention mechanisms might be needed.

In addition to syllabus statements on academic misconduct, several researchers examined whether academic dishonesty warning statements can be effective in cheating reduction. Beck (2014) utilized a cheating warning statement as a security mechanism during unproctored (*N*=19) and proctored (*N*=81) exams in the introductory economics university course. Right before the exam, the proctored group was warned not to cheat with a short description of consequences of academic dishonesty. The same warning was posted on the course site for the unproctored group with the addition that students' actions during the online test can be viewed by the instructor. The t-test demonstrated no significant difference in scores on proctored and unproctored exams (M_{pr} =40.21, M_{up} =40.63, *t* =.347, *p* >.05). Beck (2014) noted possible effectiveness of the cheating warnings and recommended to study them in future research.

The effectiveness of cheating warning statement, which emphasized the consequences of cheating, in comparison with honor code, a written pledge of academic honesty, was studied in the previously mentioned "honey-pot" website study conducted by Corrigan-Gibbs et al. (2015). The researchers randomly assigned all 404 participants into three groups: honor code group, warning statement group, and no honor code-warning statement group. At the beginning of the exam, the students in the first group read the honor code statement; the warning cheating statement was provided to the second group, and the third group was not exposed to the honor code or warning statements. The number of "cheaters" who visited the honey-pot website or plagiarized on the free-response question was analyzed across all groups. A Chi-square test identified that cheating was the highest in the control group (34%), followed by the honor code group (25%) and the warning statement group (15%). There was a significant difference in cheating between the control and warning statement groups (χ^2 (1,267) =11.9, p<.001); the difference between the honor code and control groups was not significant (

χ^2 (1,267) =3.74, p =.11). Corrigan-Gibbs et al. concluded that the warning statement was more effective than the honor code and could be implemented as an easy-to-use alternative to it.

The effect of popup cheating warning about the unauthorized use of the Internet during unproctored web-based exams was studied by Diedenhofen and Musch (2016) during the laboratory experiment conducted by the Department of Experimental Psychology at the University of Dusseldorf, Germany. Two hundred nighty eight participants took an unproctored basic knowledge exam. The exam had 16 randomized items, which were hard, but could be easily found on the Internet. After the 8th question, during the second half of the exam, a warning cheating statement popped up every time after the examinee looked up an answer. The PageFocus software was used to detect the number of times the participants were searching the Web. According to the obtained record, the participants needed at least three seconds to copy the question and pasted it into the search engine; if the exam page was left for less than three seconds, it was not considered cheating. The numbers of “look ups” before and after the warning statement popped up were compared. According to ANOVA, the number of “look ups” during the second half of the exam with the warning statement was significantly lower than during the first half of the exam without the statement (F(1,297)=562.45, p<.001, η_g^2=.43). Diedenhofen and Musch concluded that popup warning statement is an effective cheating prevention mechanism.

Physical proctoring is another widely-used institutional cheating prevention technique (CCCCO, 2013; Faire, 2013; WCET, 2009). However, this security mechanism can be expensive (Cluskey et al., 2012; Shute & Rahimi, 2017), inconvenient (Anderson & Gades, 2017; Ladyshewsky, 2015; Varble, 2014), or impossible for some online students who live far away from proctoring centers (O'Reilly & Creagh, 2016). Students have to spend time and effort to find an approved proctor (Anderson & Gades, 2017; Cluskey et al., 2012); their instructors spending a lot of time to coordinate the entire procedure (Anderson & Gades, 2017). Salaries of staff who administer proctoring, maintaining the centers, and potential decrease in enrollment due to loss of students who unable to come to the proctoring centers constitute the cost to the institutions (Cluskey et al., 2012). For these reasons, instructors and their institutions may avoid requiring physical proctoring for exams in online programs (Bandyopadhyay & Barnes, 2014). Thus, physical proctoring does not satisfy the demands of instructors and their students.

Bandyopadhyay and Barnes (2014) surveyed 348 US online instructors in the nationwide field study to examine which proctoring services the instructors used in their classes. About 14% of the participants used remote proctoring, 21% testing centers, and over 65% did not use any proctoring services. The professors who did not use any proctoring were asked why they did not use proctoring and how they promoted academic integrity during the test. Over 40% of these instructors were not concerned about cheating, 30% did not require proctoring to make it convenient for students, 14% said it was comfortable for instructors, and 16% said their institutional policies prohibited any form of proctoring in fully online classes. To prevent cheating without proctoring 54% of faculty reminded students about academic integrity, 29% used multiple versions of the test, 42% used randomization of exam items, 71% restricted the testing time, 55%

allowed notes, and 7.5% did nothing. Bandyopadhyay and Barnes concluded that if instructors do not use proctoring, an effective combination of other mechanisms should be incorporated to preclude cheating.

Remote or online proctoring, an alternative to physical proctoring, was introduced by Kryterion in 2006 with the full scale operations starting in 2008 (Foster & Layman, 2013). Remote proctoring can be live and performed by a human, who is monitoring an exam through a webcam, or can be done automatically by recording an exam administration (O'Reilly & Creagh, 2016). This type of exam security is at an early stage of development and has some pitfalls because of this (Anderson & Gades, 2017; O'Reilly & Creagh, 2016). The associated cost, additional time to connect, comparability with physical proctoring, and effectiveness of remote proctoring were investigated in the studies described below.

Bedford, Gregg, and Clinton (2009) conducted a two-stage case study, the main goal of which was to examine implementation of Remote Proctor, software and hardware that provides automatic biometric identification and monitoring of testing environments. The cost of the device per each student was $150 plus $30 for one-year license. Thirty-one student volunteers installed Remote Proctor on their computers, registered the installation, inputted their credentials, fingerprint and pictures, logged in to the Blackboard site, took a sample test, and completed a survey after finishing the tests. On average the installation took 30 minutes; about 48% of the participants supported adoption of Remote Proctor, 22% did not, and 30 % had no opinion. Structural equation model (SEM) indicated that the students perceived Remote Proctor as useful in reducing cheating, easy to set-up, and were willing to adopt the product (χ^2 =9.78, *df* =8, *p* =.281, *RMSEA*=.086) (Bedford et al., 2009).

During the implementation stage, several difficulties were identified: integration of Remote Proctor into Blackboard required some adjustments, faculty had to agree and learn how to develop and apply the Remote Proctor exams, the streaming video of up to 2500 exams takers required large storage capacity, which was expensive, and ongoing technical support for all faculty and students in need had to be provided (Bedford et al., 2009). In spite of the difficulties, Remote Proctor was fully implemented in all online university programs. Students with accommodations and students who could not afford paying $180 had an option to take exams at the university proctoring center with a human proctor. Bedford et al., (2009) concluded that with the further development, remote proctoring can become a widely-used tool for maintaining academic dishonesty in distance education.

Davis, Rand, and Seay (2016), in their study with 261 accounting students at Tennessee Tech University, utilized an improved version of Remote Proctor, Remote Proctor Now (RPN), and slightly different, but still multiple-steps procedure. The university where the study took place had a paid licensing agreement with Software Secure Inc. The participating instructor, who taught three different accounting courses, registered the final exam for each class with Software Secure Inc. website, providing the exam dates and information about permitted resources. The students logged in to the Blackboard, clicked on the RPN link created by the instructor, completed a system test for audio and video requirements, paid a fee of $15 online, downloaded and installed the

RPN software, completed an identity verification, showed using their webcams the desks and rooms' surroundings, and were automatically redirected to the exam administered on the publisher's website. The students paid the fee and downloaded the software every time they had to take a test with RPN. During this study, Software Secure personnel identified one cheating incident when a student searched the internet to find the solution to a problem (Davis et al., 2016).

All three classes took the final exam online: one (N=60) with the RPN, the second one (N=112) without the PRN and the third one (N=89) in the computer lab with a human proctor (Davis et al., 2016). The OLS model showed that both variables *Proctored with RPN* and *Proctored with human proctor* had a negative significant effect on the final exam scores ($Coeff_{RPN}$ =-.152, t_{RPN} =-6.26, $Coeff_{HP}$ =-.0895, t_{HP} =-3.55, F =8.48, p <.01, R^2 =0.237). Moreover, the final exam scores were significantly lower for RPN group in comparison with a human proctor group (M_{RPN} =.671, M_{HP} =.756, M_{NP} =.825, F =4.40, p <.05). Davis et al. (2016) concluded that remote proctoring has the potential to preclude academic dishonesty during online exams.

Bedford et al. (2009) and Davis's et al. (2016) studies indicated that online proctoring can be effective in reducing cheating. However, the same studies demonstrated that utilization of remote proctoring is costly, time consuming, and requires multiple steps for the installation. Many community college students come from low income families and cannot afford paying an additional fee for each exam. According to The Institute for College Access and Success (2016), in California, over 61% community college students come from such families. Moreover, enrollment in many community colleges is smaller than in big universities, which may result in more expensive school licenses or inability to purchase some services. There is a need to find an alternative to physical proctoring suitable for community college students. The given investigation was designed to fulfill this need.

Institutional Cheating Prevention Techniques Related to the Study

In compliance with the Higher Education Opportunity Act of 2008, public law 110-315, each student in California Community College system, including the college where the study took place, is provided with a unique login and password (CCCCO, 2013), which allow for accessing the school LMS course websites. All online students must show their valid IDs before each proctored web-based exam. The college, which has a campus-wide policy on academic misconduct, requires each instructor to include a cheating statement on the syllabus. The cheating statement on the syllabus for the introductory statistics course, in addition to the school policy, includes the detailed description of what constitutes academic dishonesty during unproctored web-based exams with clearly stated consequences of cheating. Honor Code is not used in the department because the college did not adopt it, and because several recent studies found it less effective than cheating warning statements (Corrigan-Gibbs et al. 2015; Diedenhofen & Musch, 2016). The warning statement with the repercussions is emailed to the students before each unproctored web-based exam. A short version of the warning statement is posted on the course website and on each web-based exam page. Physical proctoring was used during the proctored web-based exams involved in the study. Remote proctoring was not considered as an alternative to physical proctoring because it is not institutionalized by the college, and an individual use of the service is not suitable for community college students due to the high cost, big installation time, and cumbersome procedure. Instead, the department focused on an alternative to physical proctoring based on the systematically selected nonbiometric security mechanisms available in most currently used LMSs.

Studies Related to Effective LMS Security Mechanisms

Daniel and Broida (2004) examined the effect of weekly web-based quizzes on chapter exams' performance in face-to-face psychology university course. Three sections of the course were involved in the study: the first one (*N*=44) did not have weekly quizzes (NQ); pencil-and-paper weekly quizzes (Q) were administered in the second section (*N*=42). The same quizzes, but in web-based format (WBQ) with automatically scored multiple-choice and short answer questions, were given in the third section (*N*=39) during 24 hours before class. All three sections were taught by the same instructor who used the same materials and exams in all classes. When the analysis of the chapter exam scores in the middle of the semester demonstrated that the students in WBQ group did not do significantly better that students with no weekly quizzes ($M_{\textit{no quizzes}}$ = 48.95, $M_{\textit{web-based quizzes}}$ = 49.75, $M_{\textit{in class quizzes}}$ = 59.45, t (81) = .64, p > .05), the researcher added randomization of the web-based quizzes' items and reduced testing time to preclude cheating and force students to study better. The revised web-based quizzes were administered during the second half of the semester. The researchers applied a one-way ANOVA and found a significant effect (F (2, 122) = 81.70, p < .01, η^2 =.57). According to Bonferroni post hoc test, there was a significant difference between the NQ group and

WBQ group (*t* (81) = 11.47, *p* < .001) and between the Q group and NQ group (*t* (84) = 10.45, *p* < .001), and no significant difference between the Q group and WBQ group (*t* (79) = .54, *p* > .05). Daniel and Broida concluded that, with randomization of questions and reduced testing time, the students cheated less, studied more, and performed better on chapters' exams.

Similar to Daniel and Broida (2004), in the study with economics university students, Beck (2014) utilized randomization of exam items, to reduce collaboration, and restricted time, which did not allow for searching the Internet and other resources. However, she also added one question per-page and blocked backtracking to minimize chances of using outside help, and feedback was not provided until all students took the test to eliminate dissemination of answers (Beck, 2014). The warning statement was posted on D2L LMS website (V. Beck, personal communication, May 17, 2015). The analysis of the students' scores on a pencil-and-paper proctored exam (*N* =81) and on the same unproctored web-based exam (*N* =19), with the security mechanisms described above, showed no significant difference (M_{pr} =40.21, M_{up} =40.63, *t* =.347, *p* >.05). The GPA and other human capital variables were not significantly different across the groups (*p* <.05). Beck (2014) concluded that the combination of the security mechanism used in the study reduced cheating to the level comparable with cheating during proctored exams.

Stack (2015) compared criminology university students' performance during a proctored pencil-and-paper multiple-choice final exam and the same unproctored web-based exam. In addition to the security mechanisms used by Beck (2014), Stack utilized the Blackboard lockdown browser, to block emailing, web searching, and copying the exam questions, and synchronous testing, which did not allow for taking the exam at one time and help a friend to take the same exam at different time. Synchronous testing also minimized dissemination of the exam's content among the students while the exam window was open. The students could not access the exam before and after scheduled time. The researcher utilized the regression analysis and found no significant difference in student performance on proctored and unproctored final exams (R^2 =. 343, b = 1.08, p > .05). Stack concluded that the combination of the cheating reduction techniques used in the investigation had a security level comparable with proctoring.

In the study with postgraduate business students, Ladyshewsky (2015) also used randomization of exam questions, restricted time, and blocked backtracking. However, unlike Stack (2015), Ladyshewsky did not use lockdown browser. Instead, the researcher utilized the higher levels of Bloom's taxonomy understand, analyze, and apply exam questions as a security mechanism, which could substitute for lockdown browser: Even if the students would try to search the Internet, the needed information could not be found there (Ladyshewsky, 2015). One section of a management and leadership course was offered each semester. A 25-item multiple-choice test was developed and administered during the nine consecutive semesters. Six out of the nine sections were taught by the same instructor: in three sections, the students (N =98) took the proctored pencil-and-paper exam; in other three (N =63), the unproctored web-based version of the exam was administered during a four-day period. An ANOVA for the six sections taught by the same instructor indicated a significant difference in scores (F (5, 155) = 2.612, p = .027) (Ladyshevski, 2015); however, the researcher did not investigate whether there was a significant difference between the scores on unproctored and proctored exams (Mp_1 =71.75, Mp_2 =76.12, Mp_3 =78.25, Mup_1 =70.96, Mup_2 =73.82, Mup_3 =72.44) because he focused on change in scores over time (R. Ladyshewsky, personal communication, July 29, 2016). Ladyshevski (2015) suggested that, because the mean of scores for the unproctored exams was lower than for the proctored one (*Mp* =75.4, *Mup* =72.4), the higher order thinking exam items in a combination with randomization and blocked backtracking were effective.

Higher order thinking exam questions were also incorporated in Corrigan-Gibbs' et al. (2015) “honey-pot” website study. Additionally, the researchers used multiple versions of the same exam, randomizations of question, and restricted time as cheating reduction mechanisms (Corrigan-Gibbs et al., 2015). The unproctored web-based final exam was administered to 404 individuals. The exams scores of the students who visited the honey-pot website and plagiarized *(N* =100) were significantly lower than the scores of other test takers ($M_{cheaters}$ =30% correct, $M_{not\ cheaters}$ 40% correct, t = 4.18, p <.0001). Corrigan-Gibbs et al. (2015) concluded that exam questions focused on critical thinking, different versions of the exam, and randomizations of the test items were effective mechanisms that increased credibility of the exam: the students who tried to cheat could not obtain unfair better scores.

The Study's LMS Security Mechanisms

When the math department of the college where the investigation took place decided to implement web-based testing in the introductory statistics course curriculum, a combination of particular mechanisms was selected based on thorough analysis of the cheating strategies and methods of their prevention described in the best practices and the literature. The choice of the cheating prevention techniques also depended on the availability of the mechanisms in the college LMS Moodle, college and department policies, structure of the course offerings, and student culture of the college. The college offers multiple sections of the course every semester scheduled daily from early morning to late evening. The students in different sections of the course know each other well because they often take other courses together. For this reason, to prevent dissemination of the exam questions and reduce friend-based cheating and unauthorized collaboration, the department decided to administer all unproctored exams synchronously and make them not visible before and after the scheduled time. This approach was used in the previously described Stack's (2015) study. Blocked backtracking and one question at a time used by Beck (2014) were incorporated to reduce the opportunities for getting outside help and copying the exam items. Deferred feedback and hidden marks were chosen to prevent the sharing of answers (Beck, 2014; Stack, 2015); randomization of exam items aimed to reduce collusions (Beck, 2014; Corrigan-Gibbs et al., 2015; Daniel & Broida 2004; Ladyshewsky, 2015; Stack, 2015) and different versions of the test were developed to decrease unpermitted group work (Corrigan-Gibbs et al., 2015). The department did not utilize lockdown browser because it was not available in the college LMS. Moreover, with the recent dissemination of cell phones and other portable wireless devices, the lockdown browser is not very effective in limiting students' opportunity to search the internet (Paullet, Douglas, & Chawdhry, 2015). Instead, the department implemented higher order thinking exam questions suggested by Ladyshewsky (2015) and Corrigan-Gibbs (2015). Restricted duration of the exams was incorporated to increase the effectiveness of all other cheating prevention mechanisms (Beck, 2014; Corrigan-Gibbs et al., 2015; Daniel & Broida 2004; Ladyshewsky, 2015; Stack, 2015). The given study investigated whether this particular combination of security mechanisms is an effective alternative to proctoring by utilizing appropriate methodology and methods.

Research Related to the Study's Methodology and Methods

Quantitative strategy of inquiry, which numerically measures phenomena under investigations through systematic statistical analyses of relationships between involved variables (Creswell, 2013; Sukamolson, 2010), was chosen for the given study. This choice fits the main goal of the investigation, which is to analyze a relationship between web-based exam format, proctored vs unproctored, and exam scores collected in numerical form. The students, whose archived exams scores were analyzed in the given study, were not randomly assigned to either their classes or exams that were part of a regular educational practice. For this reason, a randomized experiment was not suitable for the given investigation. Instead, a quasi-experiment, which does not employ randomization, but otherwise possesses attributes similar to randomized experiments (Kim & Steiner, 2016; Shadish et al., 2002) was utilized.

Since Campbell and Stanley (1963) introduced the theory of quasi-experiments, this type of research methodology has become widespread in natural settings where random assignments into groups are either impossible or unethical (Cook & Campbell, 1979; Kim & Steiner, 2016; Shadish et al., 2002; Trochim, 1986). However, lack of randomization in quasi-experiments may decrease their internal validity and question trustworthiness of their causal inferences (Kim & Steiner, 2016; Shadish et al., 2002; Shadish, 2011). For these reasons, several researchers have been developing theory and practice of quasi-experiments with emphasis on improving nonrandomized experiments' validity to a level comparable to the validity of randomized experiments (Campbell & Stanley, 1963; Kim & Steiner, 2016; Shadish et al., 2002; Shadish, 2011; Trochim, 1986). Campbell and Stanley (1963) and Cook and Campbell (1979) introduced several quasi-experimental designs, discussed their strengths and weaknesses with respect to valid causal inferences, and described some suitable statistical analyses. Shadish et al. (2002) continued Campbell and Stanley (1963) and Cook and Campbell's (1979) work, focusing on generalization, emphasizing the importance of the design elements in reducing threats to validity, and prioritizing the magnitude of the effect rather than its significance.

According to the theory of experiments, the major advantage of random assignment into groups is minimization of the selection bias: randomly assigned units do not differ initially on any measured and unmeasured variables (Campbell & Stanley, 1963; Cook & Campbell, 1979, Shadish et al., 2002). Quasi-experiments do not employ randomization; for this reason, selection bias is the main threat to internal validity in most quasi-experiments (Shadish et al., 2002). Selection bias and other threats to validity of a particular quasi-experiment should be carefully identified, minimized or ruled out during the design stage by adding special design elements such that none or just a few biases are left for reduction through statistical analyses (Shadish et al., 2002). Shadish et al. (2002)

hypothesized that quasi-experiments, the designs of which eliminate or substantially reduce selection biases and rule out most other threats to internal and external validity, have the potential to produce causal estimates comparable with causal inferences of randomized experiments. To test Shadish's et al. (2002) suppositions empirically, several researchers examined the conditions under which the selection bias can be reduced and nonrandomized experiments may give results comparable with the results of randomized experiments (Pohl, Steiner, Eisermann, Soellner, & Cook, 2009; Shadish, Clark, & Steiner, 2008).

Studies Related to Reduction of Quasi-Experiments' Selection Bias

Shadish, Clark, and Steiner (2008) randomly divided university psychology students in Memphis, USA, into two groups to explore whether a randomized experiment and the corresponding quasi-experiments can yield similar results. In the first group (*N* =235), the participants were randomly assigned to mathematics (MT) (*Nm* =119) or vocabulary (VT) training (*Nv* =116). The individuals in the second group (*N* =210) self-selected the type of training (*Nm* =79, *Nv* =131) and attended the same 15-minute training sessions as the students in the randomized group. All students were pretested on 25 covariates including SAT, GPA, math and vocabulary pretest scores, previous math experience, personality type, and took the math and vocabulary posttests. To reduce the selection bias, all students were from the same university, had the same majors, grade level and age. The participants in self-selected group were similar except individuals who chose the VT had math anxiety and lower motivation than students who selected the MT (Shadish et al., 2008).

Shadish et al. (2008) found that on average the students in the randomized MT group performed 4.01 out of 18 points higher on the math posttest than individuals in the VT (Shadish et al., 2008). The participants in the quasi-experimental group who received MT earned 5.01 points more on the same math posttest than the students who selected to do the VT. The researchers used Δ=|4.01-5.01|=1 as a measure of the bias in the nonrandomized group in comparison with the randomized one; Δ=0 indicated absence of biases. The participants in the randomized group who went through the VT earned 8.25 out of 30 points more than the students who took the MT. The individuals in the nonrandomized VT group earned 9 points more on the same vocabulary posttest than the students in the nonrandomized MT group with Δ=|8.25-9|=0.75. There was only borderline evidence of significant differences in the results of random and nonrandom groups (*Mdiff rand math* =4.01, *Mdiff nonrand math* =5.01; *Mdiff rand voc* =8.25, *Mdiff nonrand voc* =9, *p*-values were not provided). The researchers concluded that the findings of their study suggested that the results from nonrandomized experiments with minimal differences between treatment and control groups can be similar to the randomized experiments' results. Shadish et al. (2008) recommended conducting further studies on this topic in different settings.

Pohl et al. (2009) replicated Shadish's et al. (2008) experiment by randomly dividing 202 psychology students at the Berlin University in Germany into two groups. The first group (*N* =99) had random assignments in math (N_m =55) and English (N_E =44) trainings; the students in the second group (*N* =103) self-selected which training to attend (N_m =55, N_E =48) (Pohl et al., 2009). The math training and math pretest and posttest were translated from Shadish's et al. (2008) materials. The English training in Berlin was focused not on vocabulary, but on basic grammar. The participants were pretested on math and English and 25 covariates identical to the covariates used by Shadish's et al. (2008). Similar to Shadish's (2008) study, the groups in Pohl's et al. (2009) investigation were not identical, but alike in many preexisted characteristics and were treated equally except the type of assignment into groups (Pohl et al., 2009).

Pohl et al. (2009) found that the biases in the nonrandomized group in comparison with the randomized group were Δ=|14.2-14.5| =0.3 on the math posttest and Δ=|-1.5-(-8.8)|=7.3 on the English posttest (Pohl et al., 2009). Thus, the direction and size of the self-selected bias in the English treatment group was opposite and larger in Berlin perhaps because the German participants, nonnative English speakers, selected the English training to improve their skills, while American students selected the vocabulary training mostly because they did not like math (Pohl et al., 2009). The participants who took the math training in the randomized and nonrandomized groups did not perform statistically differently (*Mdiff randmath* =14.2, *Mdiff nonrandmath* =14.5, *p* was not provided); the students who participated in the English training group had significantly different scores (*Mdiff randEng l*=-1.5, *Mdiff nonrandEngl* =-8.8, *p* was not provided). Pohl et al. (2009) concluded that, with the control for most differences between the groups, it is possible to model quasi-experiments with unbiased treatment effect.

The findings of the studies described above showed that quasi-experiments may have accurate effect estimates if their designs include detailed and reliable criteria by which individuals are divided into conditions, if their groups are selected from the same location and have similar characteristics, and if they utilize large sample size (Pohl et al., 2009, Shadish et al. 2008, Shadish, 2011). High quality measurement of selection into groups is important for clear identification of threats to internal validity. This criterion is especially crucial in nonrandomized experiments that utilize archived data (Shadish, 2011). Similar individual characteristics of the participants in treatment and control groups allow for the reduction of the selection bias in the design stage so that researchers do not need to remove biases during statistical analyses. The statistical selection bias adjustments require large sample size: over 200 participants in each experimental and control group (Shadish, 2011).

In addition to the selection threat, quasi-experiments with more than one assessment involved may possess testing effect bias (Shadish, 2002), which is also known as retesting effect (Villado, Randal, & Zimmer, 2016). Testing effect occurs when participants of a randomized or nonrandomized experiment are tested more than once and this repeated testing contaminates the treatment effect (Cook and Campbell, 1979; Hausknecht, Halpert, Di Paolo, Moriarty-Gerrard, 2006; Song & Ward, 2015; Shadish, 2002). This threat to internal validity can be especially strong in a within-subject study where each participant goes through all conditions and takes all tests included in the experimental design (Shadish, 2002). There are two major types of testing effect: practice effect and fatigue effect. The practice effect leads to better performance due to learning occurring during a test rather than manifestation of the independent variable; the fatigue effect results in lower performance not because of ineffectiveness of the intervention, but due to tiredness or boredom from taking the same test multiple times (Cook and

Campbell, 1979; Shadish, 2002). Several researchers examined whether practice and fatigue effects influence participants' performance and how these threats can be reduced through design elements.

Studies Related to Retesting Effect and its Reduction

One of the first studies on possible impact of retesting on performance was conducted by Spitzer (1939) with 3,605 students who constituted the entire population of six-graders in Iowa. The participants were arbitrarily divided into 10 groups. Each group took the same pretest on reading retention, the mean scores of which varied from 15.00 to 15.04. These means were very close to the population mean (not stated in the article), which demonstrated that each group represented the population well with respect to reading retention (Spitzer, 1939). The students in each group read the same passage and took the same 25-item multiple-choice test on the reading content. To measure whether retention of the reading changed and how fast, Spitzer (1939) administered the test and corresponding retests at different schedules across the 10 groups during the next 63 days. The students in all groups did not get any feedback between the tests; the number of retests varied from none to two (Spitzer, 1939).

One of the 10 groups took the initial test right after the reading, the first retest on the second day, and the second retest on the 21st day (Spitzer, 1939). The mean scores were gradually decreasing (M_{test} = 13.23, $M_{retest\ 1}$=13.07, $M_{retest\ 2}$ = 12.18) (Spitzer, 1939). The author did not discuss whether the difference between the scores within each group was significant. Another group also took the test right after the reading, but the first retest was administered in one week, and the second retests in 63 days since the initial reading (Spitzer, 1939). The means in this group were slightly lower and decreased more rapidly (M_{test} =13.20, $M_{retest\ 1}$ =11.84, $M_{retest\ 2}$ = 10.74). The students in the third group took the first test on the next day after the initial reading and had only one retest in two weeks. The means in this group were much lower than in the previous two groups (M_{test} = 9.56, M_{retest} =8.93). Another group had the initial test on the 63rd day after the reading (M_{test} = 2.71), but no retest. The students in the next group took the test and immediate retest on the same day after exposure to the passage: the mean score increased only by .03 of a score (Spitzer, 1939).

Spitzer (1936) concluded that the immediate retest, and retests at other time intervals did not demonstrate practice effects, but the students retained more when the initial test was administered sooner after the exposure to the passage. Because the scores decreased in all test-retests configurations, the researcher inferred that performance went down not because of fatigue, but because of forgetting (Spitzer, 1936). On average, the students forgot from 26% to 28% in one day after the reading, and from 39% to 44% in one week. By comparing the percent of participants who answered each test question correctly, the researcher found that there was no relationship between the rate of forgetting and item difficulty (Spitzer, 1939). Spitzer (1939) recommended using his findings in developing adequate testing schedule with repeated exams. The department where the given study took place followed Spitzer's (1939) recommendation.

Similar to Spitzer (1939), Catron (1978) investigated whether practice effect occurs if individuals are retested immediately after the initial test. However, unlike Spitzer (1939) whose study was limited to six-graders and reading retention tests, Catron

(1978) retested university students on their IQs. Thirty-five male introductory psychology students took two identical Wechsler Adult Intelligence Scale (WAIS) tests as a part of their course requirements at Wake Forest University. None of the participants had taken the test previously. WAIS, a commonly used IQ test developed in 1955, measured Verbal IQ, Performance IQ, and combined Full-Scale IQ. The Verbal part constituted of Information, Comprehension, Arithmetic, Similarities, Digit Span, and Vocabulary sections; the Performance scale included Digit Symbol, Picture Completion, Block Design, Picture Arrangement, and Object Assembly. The second administration of WAIS occurred in five minutes after the first one; the students were not aware about the second attempt, and no feedback was given between the tests (Catron, 1978).

Unlike Spitzer’s (1936) study, the retest’s scores in Catron’s (1978) investigation were significantly higher: Verbal (M_{test} =122.3, M_{retest} =125.4, *Mdiff* =3.1, *t* =4.36, *p* < .01), Performance (M_{test} =114.7, M_{retest} =128.9, *Mdiff* =14.2, *t* =14.59, *p* < .01), and Full Scales (M_{test} =120.3, M_{retest} =128.6, *Mdiff* =8.3, *t* =13.01, *p* < .01) (Catron, 1978). All mean differences in the Performance subset were significant (all *p* < .01). In the Verbal part, the students’ scores differed significantly only in Arithmetic (*Mdiff* =1.48, *t* =4.66, *p* < .01) and Comprehension (*Mdiff* =1.48, *t* =4.66, *p*< .01). The smallest differences in the Verbal scale were observed in Vocabulary (*Mdiff* = .03, *t* = .23, *p* > .01). All 35 participants demonstrated higher Performance and Full-Scale IQ. However, only 24 out of 35 students gained in Vocabulary IQ: three individuals had the same score and eight lost from one to four points. On average, the students completed the second test 23 minutes (32%) faster than the first one (M_{test} =72.6 min, M_{retest} =49.6 min, *Mdiff* =23 min, $p < .01$) (Catron, 1978).

The participants of the study could not look up answers or share content of the test items with friends during the five-minute break between the test administrations (Catron, 1978). Additionally, the average time needed to finish the test decreased significantly. Carlton (1978) concluded that the practice effect took place. The eight students who lost points in Vocabulary part during the retest could experience fatigue or boredom effect, but this effect did not lower their overall performance (Catron, 1978). Similar to Spitzer (1936), Carlton (1978) raised a question whether the practice effect changes when the time interval between the tests increases. He answered this question in the follow-up study conducted in 1979.

Unlike Catron's (1978) first study where random sampling was not utilized, in the follow-up study Catron & Thompson (1979) randomly selected 76 male university students from all introductory psychology sections and randomly assigned them to four groups (Catron & Thompson, 1979). For course credit, the students in all groups took and retook a test identical to the WAIS test used in the first study. The participants in the first group retook the test in 1 week, the participants in the second group in 1 month, the participants in the third group in two months, and the students in the last group in four months. No feedback was given between the test administrations in all groups to avoid memorization of the correct responses and reduce the practice effects (Catron & Thompson, 1979).

Catron and Thompson (1979) used a one-way ANOVA to test the initial equivalency of the four groups and found no statistical difference in IQs (the *p*-values were not stated). Overall the students in all groups did significantly better on the retest ($M_{test\ 1\ week}$ =117.84, $M_{retes\ 1\ week}$ =125.84, $M_{test\ 1m}$ =118.00, $M_{retes\ 1m}$ =123.68, $M_{test\ 2m}$ =118.95, $M_{retes\ 2m}$ =124.37, $M_{test\ 4m}$ =118.95, $M_{retes\ 4ms}$ =123.16, all *ps* < .01). However, the gain score, the difference in the mean scores between the first and second attempts, decreased gradually with time ($Mdiff_{1week}$ =8.00, $Mdiff_{f1m}$ =5.68, $Mdiff_{2m}$ =5.42, $Mdiff_{4\ m}$ =4.21 (Catron & Thompson, 1979*)*. Moreover, similar to Catron's (1978) study, the testing effect was different on Verbal IQ and Performance IQ sections. During the initial testing, all groups performed significantly better in Verbal than in Performance (*p* < .01), while on all retests the Performance scores were significantly higher than the Verbal scores (*p* < .01) (Catron & Thompson, 1979). In the last group, the Verbal IQ gain during the test-retest interval of fourth months was not significant ($Mdiff_{verbal\ 4m}$= .85, *p* < .01). Unlike Catron's (1978) article, the information on individual students whose IQ scores went down was not provided in Catron and Thompson's (1979) publication. However, the researchers noted that it is not likely for students to have fatigue effects on retests with large time intervals; instead, learning that occurs between tests may influence the retests' scores (Catron & Thompson, 1979). Catron and Thompson's (1979) concluded that longer time intervals between the test and retest do not eliminate the practice effect on WAIS test, but reduce it. Although the practice effects can manifest differently on other tests distinct from WAIS, it is important to take into consideration this threat to internal validity in all interventional studies with multiple exposures to the same tests

(Catron & Thompson, 1979*)*. Catron and Thompson's (1979) recommended finding ways to reduce practice effects in future research. Benedict and Zgaljardic (1998), whose study is discussed below, investigated one of these ways.

While Spitzer (1936), Catron (1978) and Catron and Thompson (1979) utilized exactly the same test for their multiple assessments, Benedict and Zgaljardic (1998) examined whether the use of alternative forms of the initial test, sometimes called parallel forms (Feinberg et al., 2015), can reduce the practice effect. Thirty healthy older adults, who were found through newspaper advertisement, participated in four testing sessions with 14-day intervals between the test administrations (Benedict & Zgaljardic, 1998). The 14-day interval was chosen to accommodate the laboratory schedule. Each 1-hour testing session included the revised Hopkins Verbal Learning Test (HVLT-R) and the revised Brief Visuospatial Memory Test (BVMT-R). After the participants took the first test, they were pseudo-randomly assigned into two groups: the Alternative Forms (AF) group (*N*=15) and the Same Forms (SF) group (*N* = 15*)* (Benedict & Zgaljardic, 1998).

There were four trials in which the alternative forms were administered in different order. In trial 1 (N_{AF} =4), the participants took the tests in the following order: form 1, form 2, form 3, and form 4. During trial 2 (N_{AF} =4) the order of the forms was form 2, form 3, form 4, and form 1. In the third trial (N_{AF} =3), the researchers administered form 3, form 4, form 1, form 2. During trial 4 (N_{AF} =4), the distribution of the forms was form 4, form 1, form 2, and form 3. All students in SF group took form 3 (Benedict & Zgaljardic, 1998). The researchers did not explain why the order of the alternative forms in the trials was different, and why there were four trials. The small sample size was also not discussed.

The groups were matched on individual characteristics age, education, IQ, and mini-mental state exam score (Benedict & Zgaljardic, 1998). Univariate ANOVA showed no significant difference between the groups on the individual characteristics. All alternative forms were equivalent to the first test with respect to difficulty and structure. To compare the participants' performance, the researchers used mixed ANOVA with testing sessions as the within-subject factor and groups as the between-subject factor. The researchers performed repeated measures ANOVAs within each group to examine distinctions between trials (Benedict & Zgaljardic, 1998). Similar to Benedict & Zgaljardic's (1998) study, the given investigation used univariate ANOVA to assess equality of groups with respect to instructors and course delivery mode, repeated ANOVAs to examine effects of within-subject variables *format* and *order* on *exam score*, and mixed ANOVAs to test *instructor* and *course delivery mode* effects across the test formats and interaction of the involved variables.

On the HVLT-R section of the test, Benedict and Zgaljardic (1998) found a significant *group* x *test session* interaction in trial 1 (F (3, 84) = 6.3, p < .01), trial 4 (F (3, 84) = 4.1, p < .01), and total recall (F (3, 84) = 3.41, p < .05). A significant within-subject main effect in each of these interactions was present in SF group (Fws = 19.9, p < .001), but not in AF group (Fws = .2, p was not stated). Moreover, SF group performed better than AF group in all between-subject comparison. For example, in trial 1 for AF group, $Mt1$=7.8, $Mt2$ =8.0, $Mt3$ =8.1, $Mt4$ =8.1, $Mdiff$ =.3, d=.1; in trial 1 for SF group, *Mt1*=7.9, *Mt2* =9.3, *Mt3* =10.5, *Mt4* =10.7, *Mdiff* =2.9, *d*=.9 (Benedict & Zgaljardic, 1998).

For the BVMT-R portion, *group* x *test session* interaction was significant in trial 1 (*F* (3, 84) = 6.0, *p* < .01), trial 2 (*F* (3, 84) = 3.2, *p* < .05), and total recall (*F* (3, 84) = 4.1, *p* < .01) (Benedict & Zgaljardic, 1998). Significant within-subject effects were found in both the SF and AF groups in all trials; however the effect was larger in the SF group. For instance, in trial 1, *FwsAF* = 5.1, *d* =.3, *p* < .01; *FwsSF* =32.2, *d* =1.8, *p* < .001. Similar to the HVLT-R, the mean scores of SF group in the BVMT-R increased more rapidly (*Mt1*=4.4, *Mt2* =7.9, *Mt3* =8.5, *Mt4* =8.6, *Mdiff* =4.2) than in AF group (*Mt1*=4.2, *Mt2* =5.5, *Mt3* =6.4, *Mt4* =5.3, *Mdiff* =1.7). While the testing session main effect was significant in the AF group (*F* (3, 42) = 6.6, *p*<.01), the *testing session x trial* interaction was not significant (*p* was not stated). Both the testing session main effect and *testing session x trial* interaction effect for the SF group were significant (all *p* <0.01) with the strongest effect in the first testing session (*F* (3, 42) = 16.5, *p* <.001).

Benedict and Zgaljardic (1998) concluded that the use of the alternative forms in repeated testing with the test-retest interval at two weeks significantly reduce the practice effect. Moreover, the researchers noticed that, unlike AF group, many participants in SF group spontaneously recalled items from the previous sessions, which suggested that the use of alternative forms improves construct validity of testing (Benedict & Zgaljardic, 1998). Similar to Catron (1978) and Catron and Thompson (1979), Benedict and Zgaljardic (1998) found that the practice effect was stronger in visual (BVMT-R) portion of the test than in the verbal one (HVLT-R).

The widespread integration of web-based testing forced researchers to study retesting during computerized assessments (Falleti, Collie, and Darby, 2006). Unlike Spitzer (1936), Catron (1978), Catron and Thompson (1979), and Benedict and Zgaljardic (1998) who studied the practice effects with pencil-and-paper tests, Falleti et al. (2006) examined the occurrence and magnitude of the practice effects during computerized repeated testing. Two groups of healthy university student-volunteers, aged between 18 and 40 years, participated in the study (Falleti et al., 2006). The first group (*N*=45) took the same cognitive computerized test CogState four times with 10 minutes test-retest interval and again one week later. The participants in the second group (*N*=55) were retested on the same test within 10 minutes after the first test and 1 month after that. There were no significant differences between the groups with respect to educational and IQ levels (*p* was not stated). The CogState test had eight tasks; the researchers recorded the reaction time and the percentage of correct responses (accuracy) for each task (Falleti et al., 2006).

Falleti et al. (2006) conducted a series of one-way repeated measures ANOVAs to test the presence and magnitude of the practice effect over time. In the first group, the practice effects were significant in nine of the sixteen measures (in all instances $p < .01$) with the highest change between the first and second attempts. However, the performance change, the difference in scores between the fourth and first attempts with 10 minutes test-retest interval, were significant in four tasks (*Diff task1accuracy* = .54%, *F* =6.90, *d* = .49, *p*=.01; *Diff task 5 reactime* = - 100.66 ms , *F* =11.55, *d* =- .54, *p*=.00; *Diff task 6 reactime* = - 198.75 ms , *F* =19.57, *d* =- .61, *p*=.00; *Diff task 7 accuracy* = 9.83%, *F*

=8.36, *d* = .50, *p*=.00). The performance change between the fourth and fifth attempts with one-week interval was significant in only two tasks (*Diff task 5 accuracy* =3.26%, *F* =6.81, *d* =.47, *p*=.01, *Diff task 7 reactime* = - 84.05 ms, *F* =9.93, *d* =- .36, *p*=.00). In the second group, Falleti et al. (2006) found no significant differences in performance change between the second and third attempts with one-month test-retest interval (all *p* >.01, ranged from .04 to .69). The researchers noted that, because the test performance overall did not become worse over time in both groups, the fatigue effect did not take place. Falleti's et al. (2006) results were consistent with Catron (1978), Catron and Thompson (1979), and Benedict and Zgaljardic (1998), who found that the magnitude of the practice effects decreases when the test-retest time interval increases.

While Falleti (2006) examined retesting using the same forms of the web-based CogState test, Raymond, Neustel, and Anderson (2007) compared practice effects across the same and alternative forms of a radiography computerized test. A national radiography certification test is a 200-item multiple-choice test composed of five parts: radiation protection, equipment operation, math related image protection and evaluation, radiographic procedure, and patient care (Raymond et al., 2007). Seven hundred sixty-five individuals, who did not pass the radiography exam the first time and decided to retake it, were randomly divided into two groups. The first group (*N*=663) took one of the seven versions of the alternative form during the second attempt, while the same form was administered to the second group (*N*=102). All alternative forms were equivalent to the initial test; the identical form had the same questions as the initial test, but in a different order for each student due to randomization of test items (Raymond et al., 2007). Thus, unlike Benedict and Zgaljardic (1998) who did not randomize exam items either in the SF or AF tests, the same forms in Raymond's et al. (2007) study were not 100% identical because of randomization. The SF and AF groups were similar with respect to age, gender, and test-retest interval of about 12 weeks (Raymond et al., 2007).

Raymond et al. (2007) conducted a two-way ANOVA and found a significant combined practice effect between the test administrations (*MtestSF* =70.33, *MretestSF*=73.86, *dSF* =.47, *MtestAF* =70.27, *MretestAF* =73.84, *dAF* =.48, *F*(1,763) =280.25, $p < .001$), which was consistent with other similar studies. However, unlike Benedict and Zgaljardic's (1998) study, there was no significant difference in performance between SF and AF groups (*F* (1,763) =.01, $p = .918$) in Raymond's et al. (2007) investigation. The *form x attempt* interaction was also insignificant *F*(1,763)=.01, $p = .921$) (Raymond et al., 2007). There was no significant difference in scores between five parts of the test, including math related section ($p =.145$). Raymond et al. (2007) suggested that their findings with respect to SF and AF might be different from results of previous research due to randomization of exam questions on both forms, restricted test time, and large number of items. To understand practice effects in web-based testing better, the researchers suggested replicating their investigation (Raymond et al., 2007).

Raymond's et al. (2007) study was replicated by Feinberg, Raymond, and Haist (2015) with data collected during a 200-item multiple-choice web-based medical certification exam. Three hundred thirty-eight examinees completed the test and retest within one-year interval (Feinberg et al., 2015). Two hundred nighty-five out of 338 participants took the parallel form of the exam, while 43 were randomly assigned to complete the identical form (Feinberg et al., 2015). The researchers did not discuss whether the two groups were equivalent. Feinberg et al. (2015) conducted a 2 x 2 mixed ANOVA with the first or second attempt as a within-subject factor and identical or alternative form as a between-subject factor. Additionally, the researchers investigated whether the test takers changed their answers from the first test to the second one by comparing the responses on identical forms (Feinberg et al., 2015).

Feinberg et al. (2015) found a significant combined main effect (*F* (1,336) = 108.39, *p* < .001, r =.49) meaning that on average the participants scored higher on the second attempt. Similar to Raymond's et al. (2006) study, there was no significant difference in scores across the identical and alternative forms in Feinberg's et al. (2015) investigation (*MtestSF* = 147.21, *MretestSF* =171.74, *dSF* =.47, *MtestAF* =147.42, *MretestAF* =177.19, *dAF* =.47, *F* (1,336) =.32, *ns*, *r* =.01). The *form x attempt* interaction was not significant (*F* (1,336) =1.01, *ns*, *r* =.05) (Feinberg et al., 2015). By comparing answers to multiple-choice items on both SF attempts, Feinberg et al. (2015) identified that 12.4% of answers switched from right to wrong perhaps due to guessing. Additionally, the participants spent less time on correct responses than incorrect ones (*F* (3, 24, 446) =53.44, *p* < .001, η^2 = .01) with the greatest increase in response time for incorrect-correct pattern. About 68% of the 4,924 responses were selected incorrectly on both attempts with the same letter choice. The last two facts could indicate that the participants remembered some questions and responses to them from the first attempt. Because feedback on individual test items was not provided, the examinees did not know that their answers on the first attempt were incorrect; from this point of view, such examinees can benefit from alternative forms (Feinberg et al., 2015).

Feinberg et al. (2015) noted that although there was no significant difference in the performance on identical and parallel forms in their study, the individual item analysis showed that the identical form may have a tendency to contaminate the test score. The researchers recommended using alternative forms, especially in situations where test security is important. Feinberg et al. mentioned that generalization of their findings may be limited because of the small sample size of 43 in the SF group. Also, the results may be different with test-retest interval substantially less than one year (Feinberg et al., 2015). Feinberg et al. concluded that because the participant pool was restricted to individuals who did not pass the initial test, scores on the retest could increase not because of practice effects but because the regression toward the mean (Feinberg et al., 2015).

Another study on retesting with the use of parallel and identical forms was conducted by Villado et al. (2016). Unlike Raymond et al. (2006) and Feinberg et al. (2015), who used the same and alternative forms with different individuals, in Villado's et al. study, each of the participating 307 college students took both identical and parallel forms of the Wonderlic Professional Test (WPT). The WPT measures cognitive abilities of potential employees and includes 50 questions on disarranging sentences, number series, and word problems that require mathematics and logic skills (Villado's et al., 2016). The researchers used pencil-and-paper version of the WPT (J. Villado, personal communication, January 3, 2017).

Villado et al. (2016) divided the students into four groups; each group took a different combination of two alternative forms of the WPT, Form A and Form B. The first group (*N* =81) completed Form A of the WPT during the initial testing and, in six weeks after that, the students from the same group took Form B immediately followed by Form A. The second group (*N* =73) did Form B on the initial test and, in six weeks, took Form A followed by Form B. The third group (*N* =74) initially took Form A, and then in six weeks again Form A followed by Form B. The fourth group (*N* =79) completed Form B first and after six weeks Form B immediately followed by form A. Villado et al. (2016) classified the first two groups as alternative-identical and the last two groups as identical-alternative; the assignments into groups was not random. The split-half reliability of the WPT scores across forms and attempts was .89. The researchers used test-retest interval of six weeks because this interval is usually recommended for SAT and GRE tests' retakers (Villado's et al., 2016).

Villado's et al. (2016) found no significant differences in the mean scores between the forms (*t* (305) = .78, *p* = .437, *d* =.09). Overall, the students achieved the highest score on the identical form retest (*M initial test* =30.49, *M alternative form* = 32.80, *M identical form* = 33.82) (Villado et al., 2016), which was consistent with Benedict and Zgaljardic's (1998) findings. The scores were increasing with each retest in the alternative-identical groups and the differences between the initial and each retest attempt were significant (*M initial test* =30.02, *M alternative form* = 31.73, *M identical form* = 34.33, d_{12} =.25, d_{13}=.62, *ps* < .05) (Villado et al., 2016). The retest scores in the identical-alternative groups also differed significantly from the initial test scores (*M* *initial test* =30.97, *M alternative form* = 33.31, *M identical form* = 33.87, d_{12} =.35, d_{13}=.44, all *p* < .05). However, the difference in mean scores between the second and third attempts in the alternative-identical groups were bigger than in the identical-alternative groups ($Diff_{23}$ *alternative/identical* = 2.6, $Diff_{23}$ *identical/alternate* = .56). Villado's et al. (2016) did not discuss whether the distinction between these two differences was significant, but suggested that taking two identical forms in a row could reduce the practice effect related to test-specific skills, which examinees acquire becoming familiar with the test form and structure. Villado et al. concluded that if students are provided with several practice tests so that they acquire test-specific skills before the actual exam, all subsequent tests will result in no meaningful score change due to practice effects. The studies described above suggest which combination of approaches can reduce retesting effects.

Design Elements that Reduce Retesting Effects

Longer test-retest time intervals minimize changes in retest's scores due to memorization (Catron & Thompson, 1979; Spitzer, 1936). Seven days is enough to forget a considerable amount of information reflected in test questions and answers to them (Catron & Thompson, 1979; Falleti et al., 2006; Spitzer, 1936); the interval of 1 month can eliminate the practice effects in computerized assessment entirely (Falleti et al., 2006). Larger retesting intervals also diminish fatigue effects because during these intervals individuals forget most details of the initial test, perceive a retest as a new exam, and do not get tired or bored (Catron & Thompson, 1979; Falleti et al., 2006; Spitzer, 1936). Absence of individual test items' feedback abates practice effects related to test-specific characteristics: students do not know the correct responses and cannot memorize them for retests (Catron & Thompson, 1979; Villado et al., 2016; Spitzer, 1936). Further reduction of the construct-irrelevant score changes during retesting occurs when alternative forms (Benedict & Zgaljardic, 1998) and randomization of test items (Falleti et al., 2006) are used. Practice tests administered before the actual exam may eliminate retest score's increase due to becoming familiar with the test form and structure during the initial attempt (Villado et al., 2016). Increase the number of test items on verbal task, as oppose to visual task, can reduce practice effects because verbal tasks questions demonstrated a smaller magnitude of practice effects (Benedict & Zgaljardic, 1998; Catron & Thompson, 1979; Spitzer, 1936). Additionally, Randall and Villado (2016) found that the use of security mechanisms which minimize opportunities to copy and disseminate exam questions diminish score contamination due to retesting.

Studies Related to Test Coaching and Formal Instruction

As Catron and Thompson (1979) noted, changes in student performance during repeating testing with time intervals longer than a few minutes may be related not to the practice effect per se, but to activities that take place between a test and retest. Test coaching and formal subject matter instructions are two such activities (Anastasi, 1981; Hausknecht et al., 2006; Messick & Jungeblut, 1981; Randall & Villado, 2016). Test coaching includes activities during which potential examinees become familiar with test structure, test duration, type of questions used, and test-taking strategies such as efficient pacing or elimination of unsuitable answers (Randall & Villado, 2016). Lectures, problem solving and discussions relevant to concepts covered in a test can be classified as formal instruction (Anastasi, 1981), which intended to bring desirable change score in retesting (Randall & Villado, 2016). Coaching and formal instruction effects on retest scores were investigated in a meta-analysis conducted by Hausknecht et al. (2006).

Hausknecht et al. (2006) extracted 107 independent samples (*N* =134,436) from 50 studies in which at least two attempts of a test were administered and an effect size could be calculated. Twenty-three (*N* =2,323) of the 107 samples included some test coaching; 84 (*N* =81,373) did not have coaching hours; the remaining nine studies were removed from the analysis because it was not clear whether the participants had any coaching or not. The researchers used a weighted least squares (WLS) regression to determine whether the magnitude of changes in scores on retests positively related to the number of contact hours during coaching (Hausknecht et al., 2006). The coefficient for coaching time was significant (β =.26, p < .05); the overall effect size of δ = .70 in studies with test coaching was larger than the effect size of δ = .24 in studies without any coaching; the corresponding confidence intervals did not overlap (95% *CI coaching* [.44, .83], 95% *CI no* coaching [.17, .26]). Therefore, the test coaching effect was significant (Hausknecht et al., 2006).

The examinees in 53 samples had some instruction between the initial test and retest, 54 samples did not include any instructions, and in five samples it was not clear if formal instruction took place (Hausknecht et al., 2006). By using the WLS, Hausknecht et al. (2006) found that the formal instruction coefficient was not statistically significant (β =.17, p was not stated). The researchers concluded that the formal instruction did not relate positively to the changes in scores during retesting (Hausknecht et al., 2006).

The test coaching effect can be minimized if coaching is provided for the practice tests a few weeks before the initial test and is not provided between the test-retest administrations (Villado et al., 2016). Although Hausknecht et al. (2006) found insignificant formal instruction effect, this potential threat should be taking into consideration and identified in each design with retesting (Anastasi, 1981; Randall & Villado, 2016; Shadish et al., 2002). In addition to practice, fatigue, test coaching, and formal instruction effects, changes in scores with multiple test administrations can also occur due to the order effect (Shadish et al., 2002). The order effect takes place if the order of conditions impacts the participants' behavior (Shadish et al., 2002). To assess whether, testing, coaching, or formal instruction effects confound with the interventions, the order of conditions is either randomized for each participant or counterbalanced by switching the order of treatments (Shadish et al., 2002). For example, Villado et al. (2016), in their laboratory study described above, switched the order of alternative and identical forms' administrations for each of four groups to examine whether the order in which the two forms were taken influenced change in scores. Findings and recommendations of previous researchers on how to minimize, control for, or entirely eliminate selection biases, practice, fatigue, coaching, formal instruction, and order effects are reflected in the selection of the elements of the given study's design.

The Selection of the Study's Design Elements

In 2015, the department, which provided the archived scores for the study, incorporated secured unproctored and proctored web-based exams into web-assisted face-to-face, hybrid, and fully-online introductory statistics classes. The number, order, and content of the exams and the time interval between them depended on the departmental curriculum, policies, and procedures. Some decisions of the department were informed by previous research on web-based cheating reduction mechanisms and factors that decrease possible effects of repeated testing.

The course curriculum consisted of the lecture notes, online homework, four pencil-and-paper and four web-based exams. All materials were developed by the faculty of the department, experts in the subject matter, and delivered in all sections through the college LMS Moodle. All eight exams, pencil-and-paper and online, had web-based practices tests, the format, structure, duration, and the security mechanisms of which were identical to the four web-based exams. The department developed and incorporated these practice-tests to make the students familiar with the future web-based exams and minimize rationalization and need to cheat. According to Villado et al. (2016), such practice tests also diminish the practice effect related to test taking skills. The test coaching with respect to the exam structure, duration, question type, efficient pacing and how to insert the answers took place before the first two practice tests; no additional test coaching was given after that, which rules out exam score contamination due to test coaching (Villado et al., 2016). The first secured, web-based exam was proctored in the classroom and administered in the middle of the semester. The instructors decided that the students should be very familiar with the course procedures and expectations and become accustomed to the features of the LMS by the middle of the semester and less stressed if the first web-based test is done in the classroom.

All assessments in the department are cumulative such that a few questions on each exam are similar to the questions from the previous tests. The pencil-and-paper midterm, which was used before the web-based testing incorporation, was divided into two parts: all cumulative questions moved to the online portion and all questions on the new concepts stayed in a smaller pencil-and-paper part. Thus, the next web-based exam is the alternative form of the first in class web-based exam, but administered in unproctored format in 7-10 days after the first one. The interval of 7-10 days was chosen based on Spitzer's (1936) findings about forgetting of large portion of information in 7 days after the initial reading. The students did not know that the unproctored exam is an alternative version of the in class one; no individual feedback was given between the test and retest. These two factors minimize the practice effect further (Catron & Thomson, 1979: Spitzer, 1936; Villado et al., 2016). Moreover, to reduce cheating, the exam questions were randomized; the alternative forms of the initial tests were administered. At least 7 days between the test and retest (Catron & Thomson, 1979), randomization (Falleti et al., 2006), and different theme and numerical values in the alternative version ruled out possible fatigue effects. The exams had only one question on visual task and all other were verbal-type questions, which diminished the practice effect further (Benedict & Zgaljardic, 1998; Catron, 1978**;** Catron & Thomson, 1979).

The second pair of the web-based exams was administered at the end of the last chapter in 1 month after the first pair of tests was taken by the students. Although this test-retest interval was dictated by the course curriculum, Falleti et al. (2006) found no significant practice effect on web-based exams with test-retest interval of 1 month. The second pair of the exams, Set 2, had six questions identical to the questions of the first two web-based exams in Set 1. The other 17 questions covered new but equivalent concepts at the same level of difficulty and had identical structure. The instructors decided that the first exam in the second set would be unproctored and administered outside of the class because they wanted to use all in-class time for the final exam review sessions. The alternative version of the unproctored exam in the second pair was proctored and administered in class 7-10 days after Set 2 initial test. Thus, the exams in the second set were administered in the reverse order, which allowed for examining the order effect. The elements of the study's design are represented in Figure 1.

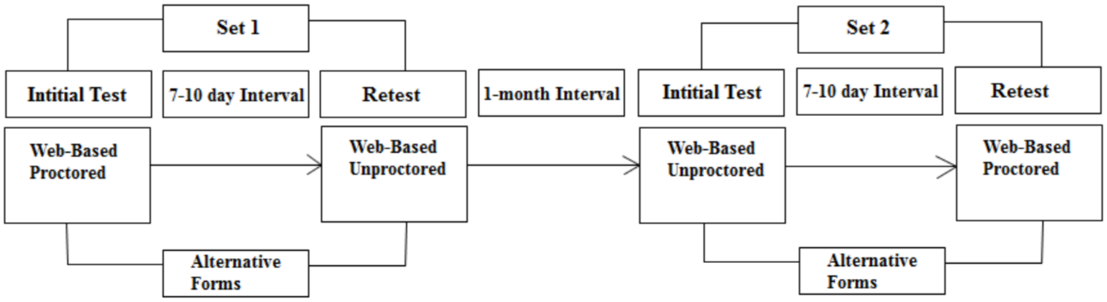


Figure 1. The elements of the study's design.

The department has been collecting data on student personal characteristics such as GPA, age, and students' scores on all assignments of all instructors who have been teaching the web-assisted face-to-face, hybrid, and fully-online sections of Introductory Statistics since Fall 2011. All these data have been used by the department for educational purposes such as creating homework and tests problems, assigning projects, and reporting student learning outcomes. The individual test scores on the four web-based exams accumulated from Fall 15 to Summer 2017 were requested. The information about exam scores, instructor, and course delivery mode was used to test the study's hypotheses. Additionally, the department provided students' GPA, age, and other student personal characteristics which were used to describe the study's population and compare the students across the groups. The institution also provided the attrition data before and after the exams were implemented, which were utilized to test whether the attrition bias due to dropout was present.

All 850 students whose exam scores were analyzed were perceived as one group of individuals who took all four web-based exams in a certain sequence. Thompson & Panacek (2006) called this design as a quasi-experimental one-group sequential design. This design can be classified as a within-subject design because in this design the same individuals participate in all conditions (Shadish et al., 2002) and each person is used as his or her own control (Thompson & Panacek, 2006). The use of a quasi-experimental one-group sequential design eliminates the selection bias and may increase statistical power: there are no preexisting individual differences between individuals within conditions (Charness, Gneezy, & Kuhn, 2012; Shadish et al., 2002; Thompson & Panacek, 2006). In addition to the absence of the selection bias, the practice, fatigue, coaching, instruction, and order effects, which are common for designs with multiple testing (Shadish et al., 2002), were controlled by the design elements described above. Thus, the given design had the high potential to produce unbiased treatment effect. Additionally, a within-subject design requires a fewer number of participants to test the same number of conditions (Charness et al., 2012; Shadish et al., 2002; Thompson & Panacek, 2006). The between-subject component of the design was added to assess the instructor and course delivery mode effects. Thus, the combination of within- and between- subject elements in the study's design allowed for reducing, eliminating, or assessing validity threats typical for quasi-experiments.

The Use of a Quasi-Experimental One-Group Sequential Design in Other Studies

A quasi-experimental one-group sequential design has been used in research on web-based testing. In 1982, the German Federal Armed Forces conducted a pilot empirical study to compare prospective recruits' performance on the first in Europe computerized adaptive Aptitude Classification Battery (ACB) test with the equivalent pencil-and-paper test (Wildgrube, 1982). The ACB had six parts: word analogy, figure reasoning, arithmetic, mechanics, spelling, and electrotechnics. A group of 208 examinees took a pencil-and-paper version of the ACB first. In some weeks or months after that, the same individuals completed a computerized version of the test, proctored at the recruiting center. Wildgrube (1982) performed the *t* test for dependent samples and found no significant difference in the participants' performance (*ps* < 0.01) in all parts except Arithmetic (*Mcomp* =12.01, *Mp&p* =11.70, *t* =1.59, *p* = .114, *r* = .79). In Wildgrube's (1982) design the participants had different test-retest interval, the retest, coaching, formal instruction, and order effects were not considered, which could bring a lot of threats to internal and external validities and contaminate the results.

Fask et al. (2015), in their study on cheating during unproctored exam in introductory statistics university course, used one group of 52 students who took two equivalent exams in the same sequence. First, the students took an unproctored web-based test, which was open for three days. The second one, a pencil-and-paper test, was administered in class on the next day after the window for the unproctored exam was closed (Fask et al., 2015). Because the researchers intended to detect cheating, no security mechanisms were used; the participants performed better on the unproctored test (*Munp* =72.96, *Mp* =65.14). Fask et al. (2015) mentioned that the students could remember the questions of the initial test during retesting, but concluded that the practice effect did not take place because students did worse on the proctored retest. However, the researchers did not discuss fatigue or order effects. The design of the given study controlled for the threats of Wildgrube (1982) and Fask's et al. (2015) investigations.

Many researchers have studied the relationship between exam format and student scores (Arnold, 2016; Beck, 2014; Fask et al., 2015; Harmon & Lambrinos, 2008; Ladyshewsky 2015, Stack, 2015; Varble, 2014). However, they used different approaches in their investigations. These approaches, their strengths, and weaknesses are described in the section that follows.

The Strengths and Weaknesses of Other Researchers' Approaches

Harmon and Lambrinos (2008) were among the first researchers who empirically studied the relationship between exam format, proctored versus unproctored, and student scores in a natural educational setting. In summer 2004, 24 online microeconomics university students took cumulative 90-minute multiple-choice unproctored web-based final exam. The unproctored exam was open for three days. In summer 2005, another group of 38 online students came to the university proctoring center to take an alternative version of the same exam also in a web-based format. The students in this group did not know about the proctored exam in advance; none of the students who took the course in

2004 took it again one year later. The same instructor taught both classes (O. Harmon, personal communication, March 5, 2015), using the same materials and assessments delivered through WebCT LMS (Harmon & Lambrinos, 2008). Both classes were allowed to use notes, books, and computer files, but not cell-phones. To prevent cheating, Harmon and Lambrinos (2008) utilized randomization of test items and alternative test forms in both sections; verbal cheating warning statement was given to the proctored group. The students who came to campus were also required to present their picture IDs.

Harmon and Lambrinos (2008) compared the academic abilities of the groups using independent *t* tests and found no significant differences in GPA, the chapter exams and final exam scores (p <.05). The researchers' main goal was to utilize the OLS analysis to assess the likelihood of cheating during an unproctored exam (Harmon & Lambrinos, 2008). Harmon and Lambrinos (2008) assumed that if the human capital variables explain smaller variation in exam score during unproctored exam than during proctored one, then cheating takes place. The R^2 -statistics of OLS models on the three chapter exams, the final exam scores, the students' GPA, age, major, and college grade level for both groups were calculated. The researchers concluded that their empirical model suggested that cheating took place during the unproctored final exam because the R^2 was smaller for the unproctored group (F_{up} =.02, $R^2{}_{up}$ =.0008; F_p =35.60, $R^2{}_p$ =.497) (Harmon & Lambrinos, 2008). Harmon and Lambrinos (2008) noted the their study was limited to two small groups of economics students, but suggested that they found some empirically supported evidence of cheating during unproctored exams and called for finding ways to improve credibility of unsupervised tests.

Harmon and Lambrinos's (2008) approach had several strengths. Both groups were treated identically with respect to the instructor (O. Harmon, personal communication, March 5, 2015), instructional materials, and assessments (Harmon & Lambrinos, 2008). Both final exams were web-based (O. Harmon, personal communication, September 25, 2016) and equivalent in relation to the content, structure, and level of difficulty (Harmon & Lambrinos, 2008). Moreover, the exams were automatically scored by the LMS (O. Harmon, personal communication, January 10, 2017), which reduced grading bias. The students in the second group did not know about the delivery mode of the final exam when they registered for the class, which could diminish the self-selection bias (Harmon & Lambrinos's, 2008). The researchers used randomization, alternative exam forms, and a warning statement to prevent cheating. They also assessed whether academic abilities of the groups were equivalent. However, there were several weaknesses in Harmon & Lambrinos's (2008) approach.

Harmon & Lambrinos (2008) did not use random assignment into groups and did not discuss how they controlled for selection biases and whether the groups represented the student population well. The unproctored exam was open for three days, which likely increased opportunities to cheat and disseminate exam questions. Although the difference in the final exams' scores was not significant (*t* = -1.32, *p* was not stated), Summer 2005 students performed better on the proctored final exam than Summer 2004 students on unproctored exam (*Mup* =73.23, *Mp* =77.15). If the academic abilities of the groups were equivalent, it is not clear why the second group had a higher mean score; the researchers did not discuss possible causes of this difference. The instructor of the courses was one of the researchers (O. Harmon, personal communication, March 5, 2015), which could bring a bias to the study (Beck, 2014). Moreover, the sample size was small, especially in the first group (*N* =24); the sample size needed for the OLS should be at least 15 individuals for each variable in the equation (Mertler & Vannatta, 2002). Harmon and Lambrinos utilized four variables in their model; therefore, they needed at least 60 students. The small sample size may raise concerns about generalizability of Harmon and Lambrinos' study's results.

Beck (2014) tested Harmon and Lambrinos' (2008) OLS model in the quantitative investigation with three sections of an economics course. The first section (*N* =19), which was delivered online, took unproctored multiple choice final exam (Beck, 2014). The unproctored exam was open for two days Thursday and Friday. The second hybrid section (*N* =21) completed a pencil-and-paper version of the same exam at the university proctoring center either on Thursday or Friday. The third section (*N* =60) took the same pencil-and-paper final exam in class on Thursday. Similar to Harmon and Lambrinos, all sections were taught by the same instructor, Beck, had the same curriculum, course materials, and assessments. However, in addition to the security mechanisms used by Harmon and Lambrinos, Beck also incorporated blocked backtracking, only one question per page, forced submission, and online cheating warning statement posted on course website for online section. All exam papers and scantrons of the hybrid group were delivered to the instructor right after the exam. Beck analyzed the data collected during one semester by combining all scores obtained on proctored exam (*N* =80) in one group and all scores during unproctored exam (***N*** =19) in another group.

According to the independent t test, there was no significant difference in exam scores between proctored and unproctored groups (Mp =40.21, Mup =40.63, t = .347, p >.05) (Beck, 2014). Beck (2014) conducted ANOVA with respect to the course delivery modes and found no significant difference in exam scores across these groups (F =.141, p =.869). The three variables GPA, major, cumulative credit were used in OLS to explain a degree of variation in exam scores: the R^2 statistics for the unproctored exam was bigger than R^2 for the proctored exam (F_{up} =18.826, R^2_{up} =.331, p <.01; F_p =13.969, R^2_p =.197, p <.01). According to Harmon and Lambrinos' (2008) logic, Beck's results demonstrated that the students were more likely to cheat during proctored exams, which, in Beck's opinion, was highly unlikely. Beck concluded that in her studies students did not do better on unproctored exam. Beck suggested that her results could differ from Harmon and Lambrinos's findings because she used more security mechanisms. Additionally, Beck questioned the completeness of Harmon and Lambrinos' model.

Beck (2014) controlled for instructor, course, and history effects. The additional security mechanisms, the online cheating warning statement with clearly described consequences of academic misconduct and blocked backtracking, were two other advantages. Beck compared student performance not only with respect to the format in which the exam was administered, proctored versus unproctored, but also across the course delivery modes, face-to-face, hybrid, online.

However, although Beck (2014) mentioned that it was not possible to utilize random assignment, she did not explain what was done to reduce selection biases. Moreover, the researcher did not discuss whether the unproctored and proctored or face-to-face, hybrid and fully-online groups were equivalent. Beck used web-based test with randomized items for unproctored group and pencil-and-paper test, the questions of which were not randomized, for the proctored group. Several researchers showed that student behavior during pencil-and-paper and web-based exams can be different with respect to performance and completion time (Bayazit & Aşkar 2012; Clariana & Wallace, 2002; Jeong, 2014; Maguire, Smith, Brallier, Palm, 2010). Similar to Harmon and Lambrinos (2008), the students in the hybrid and online sections in Beck's study could share the exam questions and solutions to them with their friends because they had 2 days to take the exam. The small sample size (*N online* =19, *N hybrid* =20) may question trustworthiness of Beck's study's results and reduce their generalizability. The researcher did not discuss either statistical power or effect size. G*Power calculator (Faul, Erdfelder, Buchner, & Lang, 2009) showed that for alpha .05, power .80, the effect size of .25, at least 53 students in each section were needed for ANOVA. Beck also mentioned that the fact that she was not only the researcher, but also the instructor could bring a bias to the study.

Similar to Beck (2014), Varble (2014) utilized security mechanisms available in the university LMS to reduce cheating during unproctored exams. However, Varble used a more systematic approach to the selection of the cheating reduction techniques. The researcher organized all known cheating prevention mechanisms into a taxonomy of cheating reduction techniques with emphasis on minimizing academic misconduct during unproctored web-based exams (Varble, 2014). Varble (2014) divided the techniques into three categories: *opportunity reduction*, *need reduction*, and *rationalization reduction* mechanisms. The researcher used mostly opportunity reduction methods to prevent cheating during unsupervised exams in his undergraduate marketing university course. Varble compared students' scores on unproctored and proctored exams and examined a relationship between Bloom's taxonomy question type and format in which exams were administered.

Varble (2014) taught one online and one face-to-face sections of the marketing course during the same semester using the same materials and tests. Both sections had weekly unsupervised web-based exams; however, the online section (*N* =28) took the unproctored web-based final exam off campus, while the face-to-face students (*N* =17) completed an equivalent pencil-and-paper version of the exam on campus during their regular class (Varble, 2014). The online unproctored and in-class proctored exams were administered on the same day; however, although the unproctored exam had limited time, it was opened during the entire day (D. Varble, personal communication, July 28, 2016). The multiple-choice final exam was randomly generated from the test bank provided by the publisher and included questions at remember, understand, analyze, apply levels of Bloom's (1964) taxonomy (Varble, 2014). The university's code of conduct stated in the syllabus was discussed with the students at the beginning of the semester. To reduce opportunities to cheat, Varble (2014) used randomization of test items, one question per page, forced test submission, restricted time, blocked backtracking, and lockdown browser.

Varble (2014) utilized mixed ANOVA to identify whether the groups performed differently on weekly unsupervised web-based tests and found a significant main effect:

the online group had higher scores than face-to-face group (F(1, 45) = 3647.63, p <.001, η^2 =.16). The researcher indicated that the assumptions of homogeneity and sphericity were met (*Both Box's M* = 146. 45, p =.304; *Mauchley's W* = .10, p =.096). To compare the students' performance on the final exam, Varble conducted the independent t test and found that the unproctored group performed significantly better on the final exam than proctored one (*Mup* = 144.93, *Mp* =112.38, t =4.47, p < .001) with an extremely large effect (d =1.38). To analyze the relationship between exam format, proctored vs unproctored, and Blooms' taxonomy question type, remember, understand, analyze, apply, Varble utilized 2 x 4 MANOVA. The researcher found that the online and face-to-face students performed significantly different across the Bloom's taxonomy levels of difficulty (F (4, 40) = 10.24, p < .001, v =.51). Four one-way ANOVAs demonstrated that a significant difference between online and face-to-face students occurred only in remembering type of questions (F (1, 43) = 31.33, p < .001, d = 1.72). To further examine the relationship between exam format and Bloom's taxonomy question type, Varble performed a discriminant function analysis. The discriminant function was significant (*Wilks' Λ* = .49, $\chi^2 = 28.91, p < .001, rc = .71$); the analysis of structured coefficients indicated that remembering contributed the most to the discrimination between the proctored and unproctored groups.

Varble (2014) suggested that if the difference in scores was due to distinct academic abilities between groups, the scores would be significantly different across all types of Bloom's taxonomy. The researcher inferred that cheating took place and unsupervised students were successful with cheating mostly on remembering questions.

Varble noted that in his study the use of security mechanisms available in the LMS was not enough to produce insignificant difference in students' performance on unproctored and proctored exams. The researcher recommended utilizing exam items that require higher order thinking skills as an additional security mechanism.

Varble's (2014) approach had several advantages. First of all, he organized methods on reducing academic misconduct into a taxonomy of cheating prevention techniques, which he used to minimize academic dishonesty during unproctored web-based exams in his courses. Similar to Beck (2014), Varble controlled for the instructor, course, and history effects. In addition to studying the effect of the exam format, proctored vs unproctored, on students' scores, Varble examined the relationship between Bloom's taxonomy question type and the exam format. The researcher used more than one statistical method to analyze the relationship: several ANOVAs and discriminant function analysis. Additionally, Varble discussed the effect sizes.

However, Varble (2014) did not link his taxonomy to Cressey's (1950) fraud triangle theory and did not discuss possible interactions between opportunity, rationalization, and need factors. The researcher did not specify whether the selection bias was present, did not compare academic abilities and other characteristics of the students in proctored and unproctored groups, and could not explain clearly why on weekly tests administered in unproctored format in both groups online students performed significantly better than face-to-face students. This significant difference could be explained by the initial individual distinctions between the groups (Shadish et al., 2002) or by the rationalization factor (Becker et al. 2006; Malgwi & Rakovski, 2008;

Tinkelman, 2012): The face-to-face students knew that they had to take proctored final exam, studied well for the weekly tests, and did not need to cheat. Similar to Beck (2014) and Harmon and Lambrinos (2008), Varble's online final exam was not synchronous, the students had one day to take the exam, which could allow for copying the test questions and sharing them with other students who were taking the exam later.

In Varble's (2014) study, the online exam was administered in a web-based mode while the proctored one in pencil-and-paper mode, which could bring additional threats (Clariana & Wallace, 2002; Jeong, 2014; Maguire et al., 2010). Although Varble utilized the lockdown browser and university code of conduct stated on the syllabus, unlike Beck, he did not use cheating warning statement posted online for the unproctored group. This could be a reason why Beck's students did not perform differently, while Varble's online students did better than face-to-face participants. Corrigan-Gibbs et al. (2015) found that cheating warning statements reduce cheating more effectively than codes of conduct. Moreover, lockdown browser may not be very effective in limiting students' opportunity to search the internet with the prevalence of cell phones and other similar devices (Paullet et al., 2015). Thus, significantly higher scores on unproctored exams in Varble's study could occur due to weak security mechanisms.

Like Varble (2014), Stack (2015) focused on reducing opportunities to cheat during unproctored web-based exams through utilization of security mechanisms available in the LMS. Lockdown browser, randomization of test items, blocked backtracking, forced completion, and limited time were the mechanisms incorporated by Stack. However, unlike Harmon and Lambrinos (2008), Beck (2014), and Varble (2014), Stack's students took the unproctored final exam synchronously during a 45-minute time frame. The students could not access the exam before and after the scheduled time (Stack, 2015). In contrast to Beck (2014) and Varble (2014) who conducted their studies during one semester, Stack (2015) analyzed exam scores collected over five years.

Ten sections of an online criminology course were taught by the same instructor, Stack, during 10 consecutive semesters (Stack, 2015). The first five sections took a proctored pencil-and-paper multiple-choice final exam on campus (N =141); the last five sections (N =170) completed an equivalent web-based final exam off campus. The same materials and assessments were used in all sections. Stack (2015) used the first course exam score as the proxy measure of the students' academic abilities.

Stack (2015) conducted a multivariate analysis with omitted exam format variable to assess trending in final exam scores over time and found that the scores neither improved nor worsened during the years of the investigation. Based on the OLS analysis with the predictors first exam score and gender, the students in the unproctored group did not perform significantly better than the students in proctored group (b = 1.08, p > .05, exact p was not stated); there was no relationship between gender and the final exam scores (b = -.876, p > .05, exact p was not stated); the model was a significant predictor of the final exam scores (F = 49.35, R^2 =.34, p < .05, exact p was not stated). Stack concluded that his findings suggested that the use of carefully selected security mechanisms can substitute for proctoring.

Stack's (2015) approach had several strengths. The researcher controlled for the instructor and course effects. Moreover, he incorporated the synchronous testing for unproctored group, which can be considered as one of the strongest cheating opportunity reduction technique (Moten et al., 2013; Shute & Rahimi, 2017; Srikanth & Asmatulu, 2014; Tinkelman, 2012). Additionally, Stack (2015) had large sample sizes (*Np* =141, *Nup* = 170). According to Tabachnick and Fidell (2007), the sample size required for a multiple regression should be at least 104+k, where k is the number of predictors. Stack used three predictors: exam format, first exam grade, and gender: therefore, the researcher had sufficient sample size to satisfy the sample size requirement. However, Stack did not explain why the variable gender was included in the analysis. The researcher also did not discuss whether the selection bias could take place. Moreover, Stack administered the proctored exam in a pencil-and-paper mode and the unproctored exam in a web-based mode, which could make the exams nonequivalent (Clariana & Wallace, 2002; Jeong, 2014; Maguire et al., 2010). Strengths and weakness of Stack's investigation were considered in the given study.

Similar to Stack (2015), Ladyshewsky (2015) analyzed proctored and unproctored exams' scores collected during several semesters. Nine sections of postgraduate management and leadership course were offered during nine consecutive terms (Ladyshewsky, 2015). The course content, materials, and assessments were the same across the sections. However, unlike Harmon & Lambrinos (2008), Beck (2014), Varble (2014) and Stack (2015), three out of nine sections in Ladyshewsky's (2015) study were taught by different instructors. To control cheating, Ladyshewsky incorporated randomization of multiple-choice exam items, limited test time, blocked backtracking, and higher levels of Blooms taxonomy exam questions. Moreover, the university where the researcher conducted his study had a very strict academic integrity policy. Ladyshewsky did not use lockdown browser, but instead purposely developed a pencil-and-paper exam questions focused on understand, analyze and apply levels of Blooms' taxonomy.

During the first four terms, 136 students took the pencil-and-paper exam (Ladyshewsky, 2015). After that, the question bank of 50 items identical to the paper-and-pencil test was created in the LMS; a set of 25 questions was randomly generated by the system from this bank for each web-based exam. During the last five terms, 114 students completed the web-based exam, which was open for a four-day period (Ladyshewsky, 2015).

Ladyshewsky (2015) utilized a one-way ANOVA including all nine tests and found that there was a significant difference in scores (Mp_1 =71.8, Mp_2 =71.3, Mp_3 =76.1, Mp_4 =78.3, Mup_5 =71.0, Mup_6 =68.3, Mup_7 =73.8, Mup_8 =72.4, Mup_9 =67.4; *F* (8, 241) = 3.628, *p* =.001). The lowest scores were in the sections taught by the different instructors during the second, sixth and ninth terms (Ladyshewsky, 2015). The researcher ran an ANOVA for the six sections taught by the same instructor and also found a significant difference in scores (*F* (5, 155) =2.612, *p* =.027). However, on average, the scores on the proctored exams were higher than on unproctored one (*Mp* =75.4, *Mup* =72.4) (Ladyshewsky, 2015). The researcher did not use any statistical methods to assess whether this difference was significant (R. Ladyshewsky, personal communication, July 29, 2016), but suggested that, due to randomization, bigger proportion of hard questions could occur during unsupervised exams, which could result in the lower scores on that exams (Ladyshewsky, 2015). The difference in students' individual characteristics between the proctored and unproctored group could be another possible cause of the lower scores during unproctored tests (Ladyshewsky, 2015). Because the participants did not do better on the unproctored exams, Ladyshewsky (2015) concluded that the higher order thinking exam questions in a combination with randomization, blocked backtracking, and strict policies on academic misconduct were effective cheating reduction mechanisms.

The utilization of higher order Bloom's taxonomy test items as a security mechanism was a distinct strength of Ladyshewsky's (2015) study. The large sample size of 250 students was another advantage. However, because the LMS test bank had more questions than the number of items randomly generated for each unproctored online exam, the pencil-and-paper and web-based exams could be not equivalent with respect to cognitive and conceptual difficulties (Ladyshewsky, 2015). Although the researcher mentioned that the students in unproctored group could be older and busier than the students in proctored group, he did not use any statistical analysis to test for differences in individual characteristics of the participants across the groups. Ladyshewsky (2015) noted that, on average, the students' scores during the unproctored exams were lower than during the proctored exams, but did not test whether this difference was statistically significant. Similarly, the researcher stated that the exam scores in the sections taught by less experienced instructors were lower, but did not assess statistically whether the students' performance differed significantly with respect to the instructors.

In contrast to all researchers discussed above, who compared exam scores of different students using between-subject design, Fask et al. (2015) utilized within-subject approach with a convenience sample of 52 introductory statistics university students. In Fask' et al. study, each student took an unproctored web-based online exam followed by an equivalent proctored pencil-and-paper test administered in-class. The researchers did not use security mechanisms because they intended to obtain an empirical evidence of cheating. Thus, both 2-hour exams were open-books, had five not randomized numerical answer problems, the blocked backtracking and lockdown brother were not activated. The unproctored exam was open for three days; the proctored exam was administered on the fourth day. The researchers used the same rubric to grade the web-based and pencil-and-paper exams (F. Englander, personal communication, October 21, 2016). On average, the student performed better on unproctored exams (*Mp* =65.14, *Mup* =72.96) (Fask et al., 2015). Similar to Beck (2014), Fask et al. questioned the completeness of Harmon and Lambrinos's (2008) model and instead used the latent variable approach. Fask et al. found that cheating had a direct effect on exam score during unproctored test and concluded that academic dishonesty took place.

The within-subject design, which minimizes selection biases in not randomized experiments (Shadish, 2011), was one of the main strengths of Fask's (2015) et al. study. Because the scores were collected during one semester and only one instructor was involved, the researchers controlled for instructor and history effects. The use of a latent variable approach in cheating detection added to the body of empirical research on academic misconduct (Fask et al., 2015). However, Fask et al. (2015) did not discuss fatigue and order effects, the validity threats typical for within-subject approaches (Shadish et al., 2002). Moreover, the proctored exams in Fask's et al. study were scored by a human, which could bring a grading bias.

Similar to Fask et al. (2015), Arnold (2016) utilized a within-subject design to study empirical evidence of cheating during unproctored web-based exams with a cohort of 461 freshmen economics university students in the Netherlands. Each student in this cohort took unproctored exams in Microeconomics, Statistics, Accounting I, and Accounting II and proctored exams in Mathematics I and Mathematics II. Summative proctored exams were administered in all courses; the unproctored exams had randomized multiple-choice questions and restricted time. The average scores on unproctored exams ranged between 7.75 and 8.34 while the scores on the proctored exams were between 4.32 and 7.49 (Arnold, 2016). Arnold (2016) applied Harmon and Lambrinos' (2008) OLS model on unproctored formative exam scores and the same subject summative proctored exam for detecting cheating and Jacob and Levitt's (2003) algorithm for determining unexpected fluctuations in test scores.

According to Harmon and Lambrinos' (2008) model, the R^2 in Arnold's (2016) study, which for the unproctored tests, ranged from .0096 to .176 and for the proctored tests from .150 and .254, indicated evidence of cheating during unproctored exams. Additionally, the correlation coefficients between Jacob and Levitt's scores in unproctored exam group were positive, ranging between .35 and .46 (p <.01), which suggested high likelihood of cheating. The correlation coefficients for proctored exams were between -.02 and .11 (p <.05), indicating low likelihood of cheating. Seven students had extremely high scores on the unproctored tests (9.5-10 out of 10 possible) and extremely low scores (1-2.3) on the corresponding summative proctored exam. Arnold concluded that the students were more likely to cheat during unproctored exams.

Similar to Varble (2014), Arnold (2016) used several statistical analysis methods to investigate the likelihood of cheating during unproctored exams, which was an advantage of his approach. Additionally, Arnold utilized a big sample size of 461 students. However, although Arnold (2016) used within-subject design, the proctored and unproctored exams scores discussed by the author were on different subjects. Thus, a student could get a higher score on an unproctored Microeconomics formative exam than on a proctored formative Mathematics I exam, not because of cheating, but rather because mathematics was harder for this student than microeconomics. Moreover, the researcher did not discuss whether formative and summative assessments on the same subject were equivalent. Also, Arnold used only two security mechanisms, randomization and restricted time, which could be not enough to reduce cheating.

Many weaknesses and strengths of the approaches described above were taken into consideration in the given investigation. Similar to Harmon and Lambrinos (2008), to avoid possible nonequivalency of exams due to differences between pencil-and-paper and web-based modes, only the scores of the exams administered in web-based mode were analyzed in the given study. To reduce self-selection bias, similar to Fask et al. (2015) and Arnold (2016), within-subject design was utilized. To control for academic dishonesty, the department, the exam scores of which were requested, utilized all security mechanisms used by Beck (2014), synchronous administration of unproctored web-based exams incorporated by Stack (2015), and higher order thinking test items suggested by Ladyshewsky (2015). Similar to Beck, I compared scores on the proctored and proctored exams across course delivery modes; like Ladyshewsky, the scores were also compared with respect to instructors.

The reverse order of proctored and unproctored exams and the test-retest interval selected by the department allowed for avoiding weaknesses of Fask et al. (2015) study by controlling for practice, fatigue, and order effects. To avoid a flaw of Varble (2014), who did not compare students between the groups, I assessed whether the students whose score were analyzed were significantly different across all groups with respect to GPA and age. Several researchers mentioned that GPA (Harmon & Lambrinos, 2008; Beck 2014; Arnold, 2016) and age (Ladyshewsky, 2015; Gallant, Binkin, & Donohue, 2015) can impact student propensity to cheat. Therefore, the comparison with respect to GPA and age allowed for more accurate interpretation of the given study's findings. Automatic grading of all exams incorporated by the department allowed for avoiding grading bias, which could be present in Fask et al. (2015) investigation. The number of the test items in the test bank created by the department was equal to the number of randomly generated questions from this bank for each web-based exam, which, unlike Ladyshewsky, preserves equivalency of the study's proctored and unproctored exams.

Justification from the Literature of the Rationale for the Selection of the Variables

The purpose of this study was to investigate whether web-based exams with systematically selected security mechanisms can substitute for proctoring. To fulfill this purpose, a relationship between format of the exam administration, proctored versus unproctored, and student scores on the exams was examined. All researchers, whose studies were discussed in the previous section, explored the relationship between exam format and score by manipulating the format in which the exams were administered, proctored vs unproctored (Arnold, 2016; Beck, 2014; Fask et al., 2015; Harmon & Lambrinos, 2008; Ladyshewsky, 2015; Stack, 2015; Varble, 2014). The work of these researchers provides justification for the rationale for the selection of the variable exam format with two levels, proctored versus unproctored, as the main independent variable, and the variable exam score as the dependent variable of the given investigation.

In addition to the format effect, Beck (2014) tested the course delivery mode effect on student performance. Beck’s explorations support the rationale for choosing the variable course delivery mode as another given investigation’s independent variable with three levels: face-to-face, hybrid, online. Moreover, Beck emphasized the importance of controlling for instructor effect. Ladyshewsky (2015), who compared students’ scores on proctored and unproctored exams across different instructors, also discussed possible different manifestation of the relationship between exam format and students’ scores across distinct instructors. Beck and Ladyshewsky’s work justifies the selection of course instructor as one more independent variable of the given investigation. Although Fask et al. (2015) did not study order effect in his research project, he noted that the order in which equivalent proctored and unproctored exams are administered can influence student performance. Fask’s et al. study supports the choice of the variable order as the last independent variable of the given investigation.

Previously mentioned Harmon and Lambrinos (2008) and Arnold (2016) applied OLS analysis and identified that GPA was a significant predictor of the likelihood of cheating (*p* <.05). Gallant et al. (2015) surveyed 1,200 undergraduate university students and found that students with low GPA (less than 2.8) and who are younger have a higher risk of being reported cheating (p <.001). These findings justify the selection of GPA and age for comparison of students across course delivery mode and instructor groups.

Many researchers have studied the variables described above (Arnold, 2016; Daniel & Broida, 2004; Beck, 2014; Fask et al., 2015; Ladyshewsky, 2015; Sivula & Robinson, 2015; Stack, 2015). The description of the findings of these researchers with respect to the study's variables begins with the analysis of what is known about the relationship between exam format, security mechanisms, and exam score.

Literature Related to what is Known about the Key Independ and Dependent Variables

The study's key independent variable, format of exam administration, proctored versus unproctored, has been examined since early adoption of web-based testing in Higher Education. In the early 2000s, Daniel and Broida (2004) studied student performance on proctored and unproctored weekly quizzes in a face-to-face psychology course at a public university in New England. Three sections of the same course taught by the same instructor with the use of the same materials and assessment participated in Daniel and Broida's study. The students in the first section (*N* = 44) did not have weekly quizzes (NQ), the students in the second section (*N* =42) completed in-class 15-minute pencil-and-paper weekly quizzes (Q), the students in the third section (*N* =39) took the same 15-minute quizzes in unproctored web-based format (WBQ). The unproctored quizzes were unsecured and opened for 24 hours. Daniel and Broida conduct a one-way ANOVA to compare the mean of Exam 1 and Exam 2 scores at the middle of the semester and found a significant difference in the means across the groups ($p < .01$, $\eta^2 =.43$). According to Bonferroni post hoc test, there was a significant difference in scores between the Q group and WBQ group ($p < .001$) and between the Q group and NQ group ($p < .001$). However, there was no significant difference between the NQ group and WBQ group ($p > .05$, exact *p* was not stated) (Daniel & Broida, 2004).

After Daniel and Broida (2004) activated randomization of test items feature, decreased the web-based quiz duration time from 15 to 7 minutes, and removed the glossary from the site, the average of the WBQ group students' scores on Exam 3 and Exam 4 administered in the second half of the semesters increased. The researchers applied a one-way ANOVA and Bonferroni post hoc test to the Exam 3 and Exam 4 scores and found no significant difference between the Q group and WBQ group ($p > .05$, exact *p* was not stated), although there was a significant difference in scores between the NQ group and WBQ group ($p < .001$) and between the Q group and NQ group ($p < .001$). The researchers concluded that the students were cheating during unsupervised quizzes when security mechanisms were absent (Daniel & Broida, 2004).

Like Daniel & Broida at the beginning of the semester (2004), Sivula and Robinson (2015) did not utilize any cheating reduction techniques during an unproctored exam in their study with graduate university students. Two sections of a research method course, face-to-face and online, were taught by the same instructors who used the same materials and tests (Sivula & Robinson, 2015). The students in the face-to-face group (*N* =20) took a two-hour comprehensive final exam with 22 short essay questions. The online group (*N* =21) completed an unsupervised web-based version of the same exam, which was open for four days, had multiple attempts, and unlimited time. The online students could use any resources, but were asked to work independently. The instructor graded both exams using the same grading rubric. Sivula and Robinson (2015) conducted a *t* test and found a significant difference in students' performance on proctored and unproctored exams (*p* =.001, *Cohen's d* =1.084); the two-sample t-interval did not include 0 (95% CI [4.04, 16.11]). The online group performed 34% better than the face-to-face group. The researchers concluded that cheating took place during the unproctored exam (Sivula & Robson, 2015).

Similar to Daniel and Broida (2004) and Sivula and Robinson (2015), Fask et al. (2015) compared students' scores on proctored and unsecured unproctored exams. Fifty-two introductory statistics university students took two equivalent exams, unproctored web-based and proctored pencil-and-paper (Fask et al., 2015). Both 2-hour open-book exams included only five numerical not multiple-choice questions. Fask et al. (2015) did not use any cheating prevention mechanisms: the questions were not randomized, the backtracking and lockdown browser were not activated, and the unproctored exam was open for three days. The students performed better on the unproctored exam (*Mp* =65.14, *Mup* =72.96). Fask et al. utilized the latent variable approach and found evidence of academic dishonesty: the path diagram connected the latent variable cheating and observed variable score on unproctored exam with the regression coefficient of 1 indicating that cheating had a direct effect on exam score during the unproctored test.

Northcutt et al. (2016) studied the use of multiple-accounts during unproctored online certificate exams administered in 115 MOOCs offered through edX platform at Harvard and MIT in 2012-2015. To detect multiple accounts' users, who create one or more "harvester" accounts to access a test's answers, copy, and paste them into the test taken through the major account, the researchers applied a Bayesian criterion detection algorithm and found that about 1,237 certificates were earned through the use of multiple accounts by 657 students. The biggest number of unfairly earned certificates (1.2%) was identified in courses, the instructors of which did not employ any cheating prevention methods (Northcutt et al., 2016).

Daniel and Broida (2004), Sivula and Robinson (2015), Fask et al. (2015), and Northcutt et al. (2016) suggested that students are likely to cheat during unproctored assessments with no security mechanisms. Moreover, scores on unsecured unproctored exams tend to be higher than on equivalent proctored exams (Fask et al., 2015; Sivula & Robinson, 2015). This knowledge inspired several researchers to utilize cheating prevention mechanisms during unsupervised we-based exams and compare student performance on secured unproctored and traditional proctored exams (Beck, 2014, Ladyshewsky, 2015; Stack, 2015). However, the findings of these researchers were mixed and sometimes controversial.

Literature Related to what is Controversial about the Key Independ and Dependent Variables

Wachenhem (2009) utilized randomization and restricted time during an unproctored final exam in a university economics course. Three sections of an economics course were taught by the same instructor who used the same materials and exams (Wachenhem, 2009). The students in the first section (*N*=9) were enrolled in Summer 2006 face-to-face class, which had a closed-book proctored web-based final exam. The students in Spring 2007 online sections (*N* = 27) took the same web-based exam also in a proctored format, while the students in Summer 2007 section (*N* =18) completed the exam in an open-book unproctored format. Wachenhem (2009) did not discuss whether she compared academic abilities and other student personal characteristics across the groups. All three sections had unproctored chapter exams with randomized items and restricted time (Wachenhem, 2009). The students who took the proctored exam were combined in one group (*N* = 36). The unproctored exam was open for 24 hours.

Wachenhem (2009) used OLS model to predict performance on the final exam based on average chapter exams' score; course delivery mode and exam format were other variables included in the model. The researcher found that the correlation between the average of chapters' exams and final exam scores was higher for the unproctored group (*b* =.917, *p* =.004) than for proctored one (*b* =.534, *p* =.004). The estimated coefficient of the variable unproctored was 13.68 while for the variable proctored was 16.52 (R^2 =.658, adj. R^2 =.638). The students in the unproctored group obtained almost one full letter grade higher on the final exam than the students in the combined proctored group (*Mp* =65.14, *Mup* =72.96). Wachenhem suggested that the higher score in the unproctored group could be explained by the fact that the text book was allowed during the final unproctored exam, but not during the proctored one. Wachenhem's results may have limited generalizability due to the small sample size.

Similar to Wachenhem (2009), Beck (2014) studied students' performance on proctored and unproctored exams in an economics university course. However, in addition to randomization of test items and restricted time used by Wachenhem, Beck also utilized blocked backtracking, one test item per page, and cheating warning statement, which included clear descriptions of the consequences of academic misconduct and was posted on the class website. In Beck's study, 19 students took a web-based unproctored exam, which was open for two days, and 81 students took an equivalent proctored pencil-and-paper version of the same test. Beck (2014) applied the independent *t* test and found no significant difference in exam scores between proctored and unproctored groups (*p* >.05, exact *p* was not stated). The researcher concluded that the results of her study could indicate that with the appropriate security mechanisms students do not do better on unproctored exams than on proctored (Beck, 2014).

In addition to Beck's (2014) security mechanisms, Varble (2014) utilized lockdown browser in his study with university marketing students. At the beginning of the term, the researcher also discussed with his students the university‘s code of conduct stated in the syllabus (Varble, 2014). During the same semester, 28 students took the secured unproctored web-based exam, which was open for 24 hours, and 17 students completed an equivalent proctored pencil-and-paper test. Varble (2014) applied the independent t-test and found that the unproctored group performed significantly better on the final exam than proctored one (p < .001, d =1.38). The researcher conducted a further analysis utilizing ANOVAs and a discriminant function analysis and determined that a significant difference between proctored and unproctored groups occurred in Bloom's remembering type of questions (p < .001, d = 1.72), which also contributed the most to the discrimination between these groups (*Wilks' Λ* = .49, $\chi^2 = 28.91$*,* $p < .001$*,* $rc =$ .71). Varble suggested using higher order thinking test questions as an additional security mechanism.

Ladyshewsky (2015) added higher order thinking items to the set of security mechanisms used by Varble (2014), but did not utilize lockdown browser. Ninety-eight postgraduate students of a research method course took the proctored pencil-and-paper exam, and 63 students took the unproctored web-based exam, which was open for four days (Ladyshewsky, 2015). Ladyshewsky (2015) conducted an ANOVA and found a significant difference between proctored and unproctored exams' scores (p = .027). However, on average, the students did better on the proctored exams than on unproctored ones (*Mp* =75.4, *Mup* =72.4). The researcher mentioned that this difference could be explained by harder questions on unproctored exams due to randomization and distinctions in personal characteristics between the groups.

Similar to Wachenhem (2009), Brallier and Palm (2015) used only randomization of test items and restricted time as security mechanisms during unproctored web-based exams in their study with undergraduate sociology students at a southwestern university. The same instructor taught one face-to-face and one fully-online section of an introductory sociology course each term during four consecutive semesters (Brallier & Palm, 2015). The same materials and assessments were used in all sections. All students took a web-based pretest which measured their baseline knowledge of the subject. During the first two semesters, the first group of students (*N* =130) took four open-book unproctored tests, including the final exam. Each unproctored web-based exam was timed, randomized, and open for 24 hours. In the last two semesters, the second group of students (*N* =116) took the same four exams, but administered in a proctored closed-book pencil-and-paper format. A cumulative test score for each student was calculated as the percent of the total points obtained on the four exams. According to the independent *t* tests, there was no significant difference in academic abilities between the groups (M_{GPAup} = 3.31, M_{GPAp} = 3.36, *t* = -.72, *p* =.48; *Mpretest up* = 52.72, *Mpretest p* = 49.91, *t* =1.73, *p* =.09) (Brallier & Palm, 2015).

Brallier and Palm (2015) conducted a 2 x 2 between-subject ANOVA with the variables exam format, proctored versus unproctored, and course delivery mode, face-to-face versus online, and found that the exam format had a significant effect on exam scores (*Mp* =68.65, *Mup* =74.66, *F* (1,242) =17.41, *p* < .001, η^2 =.07), while the course delivery mode effect was not significant (*F* (1,242) =3.45, *p* =.07, η^2 =.01). Moreover, the researchers did not find a significant interaction between test format and course delivery mode (*F* (1,242) =3.27, *p* = .07, η^2 =.01). Brallier and Palm concluded that the students performed 6% higher on unproctored exams than on proctored ones with the medium effect size of 7% and raised a concern about possible grade inflation during unproctored exams. To improve the credibility of unsupervised web-based exams, in addition to randomization and limited time, the researchers recommended using higher order thinking exam questions and synchronous testing (Brallier & Palm, 2015).

Although Stack (2015) did not utilize higher order thinking exam questions, he incorporated synchronous test administration in addition to lockdown browser, randomization of test items, blocked backtracking, forced completion, and limited time. With these security mechanisms, by using the SLO analysis, Stack found no significant difference between the scores of 141 criminology students who took the proctored pencil-and-paper final exam and 170 students who took the equivalent unproctored web-based exam ($F = 49.35$, $R^2 =.34$, $b = 1.08$, $p < .05$). The researcher concluded that a systematically selected combination of security mechanisms may result in equal performance during unsupervised and supervised exams and substitute for proctoring (Stack, 2015).

Similar to Wachenhem (2009) and Brallier and Palm (2015), Arnold (2016) used only randomization and restricted testing time as security mechanisms during unproctored exams with a cohort of 461 economics university students. By applying Harmon and Lambrinos' (2008) OLS model and Jacob and Levitt's (2003) algorithm for determining unexpected fluctuations in test scores, Arnold found that the students performed better on unproctored exams ($.010 \leq R^2 up \leq .176$, $.150 \leq R^2 p \leq .254$, $.350 \leq$ *Levitt's Score up* $\leq .460$, $p <.01$). Arnold concluded that cheating took place during unsupervised exams.

Although the researchers, whose studies are described above, obtained different results with respect to students' scores on proctored and unproctored exams, a few common themes emerged from the analysis of their investigations. In all studies where only randomization and restricted time were used as security mechanisms (Arnold, 2016; Brallier & Palm, 2015; Wachenhem, 2009), students performed better on unproctored exams. In the studies where no significant difference between students' scores on proctored and proctored exams was found (Beck, 2014; Stack, 2015), in addition to randomization and restricted time, the instructors utilized blocked backtracking (Beck, 2014; Stack, 2015), one test item at a time (Beck, 2014; Stack, 2015), cheating warning statement (Beck, 2014), and synchronous testing (Stack, 2015). Varble (2014) incorporated randomization of test items, restricted time, blocked backtracking, and lockdown browser, but his students performed better on unproctored exams. Therefore, it can be concluded that the security mechanisms' combination used by Varble was not effective enough. Ladyshewsky's (2015) students from the unproctored group performed worse than the students who took the proctored exams, but nonequivalency of the exams and participated groups could contaminate this result.

There was another commonality among the described studies. In all, except Wachenhem's (2009) investigation, the researcher used web-based mode of test administration for unproctored exams and pencil-and-paper mode for proctored exams, assuming that these two modes are equivalent. However, this assumption may be controversial because several previous researchers compared student performance on pencil-and-paper and computerized exams and obtained mixed results (Clariana & Wallace, 2002; Jeong, 2014, Maguire et al., 2010; Mayer & Krampen, 2015). The analysis of their studies is provided below.

Literature Related to Equivalency of Pencil-and-paper and Web-based Exams

Extensive research has been done on the comparison of pencil-and-paper and web-based exams mode administration since the first adoption of computerized testing in the early 1880s. Previously mentioned Wildgrube (1982) compared 208 prospective German Federal Armed Forces recruits' performance on the first in Europe computerized adaptive Aptitude Classification Battery test with the equivalent pencil-and-paper test. The researchers utilized a paired t-test and found that the participants performed better on the Arithmetic portion of the test administered in web-based mode than on the same Arithmetic portion administered in pencil-and-paper mode (*Mcomp* =12.01, *Mp&p* =11.70, *t* =1.59, *p* = .114) (Wildgrube, 1982). Mead and Drasgow (1993) examined the effect of the mode of exam administration, pencil-and-paper versus computerizes, on test scores in their meta-analysis of 29 studies conducted in 1977-1992. The researchers used a within-subject design to compute cross-mode correlation coefficients and found that timed tests were affected by the mode of test administration (R^2 =.38, *r* =.72, *p* <.01), while untimed tests were not (R^2 =.38, *r* =.97, *p* >.05) (Mead & Drasgow, 1993). Further research on this issue has been continued in a new millennium.

Clariana and Wallace (2002) compared student performance on proctored pencil-and-paper and proctored web-based multiple choice exams administered in four sections of Computer Fundamentals university course. Two of these sections (*N* =51) were randomly selected in the pencil-and-paper exam group, and other two sections (*N* =54) composed the web-based exam group. The paper-and-pencil test had 100 items with seven questions per page. The web-based test had the same items, but they were randomized, and the examinee could see only one question at a time. The backtracking option was not blocked: the students could go back and change their answers. All sections were taught by the same instructor who utilized the same materials. The final course grades were used to measure content knowledge of the participants (Clariana & Wallace, 2002).

Clariana and Wallace (2002) conducted a between-subject ANOVA and found a large significant difference in scores between the groups (*Mwb* = 83, *Mp&p* = 76.2, *F* (1,103) = 15.32, p <.001). The researchers also applied 2 x 2 ANOVA with factors *exam mode*, paper versus web-based, and *content knowledge*, low versus high, and observed a significant test mode effect (*F* (1,101) = 9.64, *p* =.002) and a significant content knowledge effect (*F* (1,101) = 12.483, *p* =.001). High-performing students outscored low-performing students on the web-based exam (Clariana & Wallace, 2002). The mode-content knowledge interaction was also significant (*F* (1,101) = 5.07, *p* =.027). Clariana and Wallace (2002) concluded that pencil-and-paper tests might not be equivalent to the corresponding web-based tests and noted that instructors and administrators must be aware of this fact.

Maguire et al. (2010) obtained results similar to Clariana and Wallace' (2002) findings, examining the test mode effect in a university accounting course offered during three consecutive semesters. The university offered two sections of the course each term. The students self-enrolled into the sections not knowing about the mode of the exam administration. All six sections of the course were taught by the same instructor and had the same three chapter tests and the final exam. Forty-three students took all four multiple-choice exams in a proctored web-based mode and 92 completed the same exams in a proctored pencil-and-paper mode. Maguire et al. did not discuss whether the questions on the web-based exams were randomized. The total scores on all four exams were analyzed (Maguire et al., 2010). The researchers applied ANOVA and found a significant test mode effect (*Mwb* = 69.77, *Mp&p* = 64.18, *p* = .0002). Maguire et al. concluded that the students who took the web-based exams performed better than the students who took the pencil-and-paper tests.

Unlike Clariana and Wallace (2002) and Maguire et al. (2010) who conducted between-subject studies, Jeong (2014) compared sixth-grade Korean students' performance on pencil and paper and web-based test using within-subject design. Seventy-three students, who were randomly selected for the study, took the pencil-and-paper test followed by the corresponding multiple choice web-based exam. The researcher did not specify the test-retest time interval, but mentioned that the order effect was not tested because the pencil-and-paper and web-based exams could be perceived by the students as different tests (Jeong, 2014). Both tests had identical four parts: Korean language, mathematics, science and social studies. The participants were familiar with computers because they received weekly one-hour computer literacy lessons beginning first grade. Jeong (2014) ran an ANOVA and found that the students performed significantly better on the Korean (*Mp&p* =82.39, *Mwb* =71.91, *F* =25.61, *p* < .01) and science (*Mp&p* =86.43, *Mwb* =81.71, *F* = 6.386, *p* < .05) sections of the pencil-and-paper exam, but there was no significant difference on the mathematics (*Mp&p* =89.04, *Mwb* =86.02, *F* = 2.077, *p* = .152) and social studies portions (*Mp&p* =85.95, *Mwb* =83.21, *F* =

1.111, p = .294). The researcher concluded that the students did better on the pencil-and-paper exam (Jeong, 2014).

Unlike Clariana and Wallace (2002), Maguire et al. (2010), and Jeong (2014) who studied the exam mode effect in natural educational settings, Mayer and Krampen (2015) conducted a laboratory experiment to examine the relationship between the exam mode, pencil-and-paper versus web-based, exam format, proctored versus unproctored, and German university students' scores on these exams. The participants were randomly divided into four groups. The first group (N = 34) took the unsupervised web-based exam, the second group (N =31) completed the same web-based exam in a proctored mode. The third group (N =43) took the unproctored pencil-and-paper version of the same test, while the last group (N =43) completed the pencil-and-paper test with a proctor. There was no difference between groups with respect to age, gender, GPA, and information literacy (p value was not provided). The exams had 35 multiple-choice items and tested students' knowledge on searching and evaluating psychology information on the Internet. Mayer and Krampen conducted 2 x 2 ANOVA and found a significant exam mode effect (Mwb = .54, $Mp\&p$ = .57, F (1,137) = 4.42, p < .05, η^2 = .031) and a marginally significant exam format effect (Mwb = .54, $Mp\&p$ = .56, F (1,137) = 2.99, p < .10, η^2 = .021). No interaction between exam mode and format was present. However, Mayer and Krampen found that the participants performed significantly better on pencil-and-paper exams than on web-based exams and slightly better on proctored exams than on unproctored.

According to the above analysis, it is known that students' scores are higher on unproctored exams when no (Sivula & Robinson, 2015; Fask et al., 2015) or just a few security mechanisms (Brallier & Palm, 2015; Wachenhem, 2009) are used, while with a systematically chosen combination of security mechanisms there is no difference in scores (Beck, 2014; Stack, 2015) or they are higher on proctored exams (Ladyshewsky, 2015). Additionally, the assumption that pencil-and-paper and corresponding web-based exams are equivalent is controversial because there are evidences that students may perform better (Clariana & Wallace, 2002; Maguire et al., 2010) or worse (Jeong, 2014; Mayer & Krampen, 2015) on web-based tests than on pencil-and-paper exams. Based on this information, a few aspects that have not been researched in the previous studies are synthesized in the section that follows.

Description of what Remains to be Studied about the Key Independ and Dependent Variables

The department, which provided the exam scores for the given investigation, incorporated randomization of test items, restricted time, blocked backtracking, synchronous testing, higher order thinking questions, and cheating warning statement posted on the course website as security mechanisms. None of the researchers whose investigations are discussed in the previous sections studied exam format and exam scores when this particular combination of the security mechanisms is used. This fact constitutes the first gap in the literature related to the key independent and dependent variables. Moreover, the researchers administered proctored exams in pencil-and-paper mode while unproctored exams in web-based mode, but, as shown above, these modes might be inequivalent. This is the second gap in the literature. Unlike these studies, the scores that were used for the analysis of the given investigation were obtained on the exams, all of which were administered by the department in a web-based mode that allows for addressing the second gap. Thus, in the given study, I addressed both gaps by analyzing the exam format effect on student scores earned on exams administered in web-based mode with the combination of the security mechanisms used by the department.

Literature Related to what is Known, Controversial, and Needs to be Studied about the Variable Order

The order in which the proctored and unproctored exams are administered is another independent variable of the study. This variable was studied by Templer and Lange (2008), who incorporated different order of the proctored and unproctored future employee's personality test in their laboratory experiment with student-volunteers at a state university in Singapore. Templer and Lange randomly divided the participants into four groups. The first group (*N* = 40) took two identical proctored web-based tests in a university lab. The second group (*N* =35) first completed the web-based test in a proctoring lab, and then took the same test online without any supervision. The third group (*N*=38), first took the online web-based test without a proctor and then completed the same web-based test in the lab with a proctor. The last group (*N* =50) took the same web-test two times in an unsupervised environment. The test-retest interval in all groups was about two weeks. The first and the fourth groups served as control and the second and third were the experimental groups (Templer & Lange, 2008).

The Internet personality battery test was provided by a company in Germany for data collection from potential Singaporean applicants for a German International Management trainee program (Temper & Lange, 2008). The test had five not cognitive sections, *Cooperation*, *Readiness to learn*, *Commitment*, *Readiness to solve problems*, *Adopting the views of others*, and one cognitive ability section, *Creativity based on rules*. The participants could have a break after each section, but Templer and Lange (2008) asked the students to take the entire battery in an unsupervised environment during one week at any location.

Templer and Lange's (2008) study design had between-subject and within-subject elements. The researchers conducted the first between-subject analysis to test whether there was a significant difference in scores between the four groups during the initial test and retest. According to ANOVA, there were significant differences across the groups in *Adopting the views of other* on the initial test (F = 2.94, p < .05, η^2=.052) and in *Cooperation* on the retest (F = 3.65, p < .05, η^2=.064). In the second between-subject analysis, the scores of all proctored tests were combined into one group, and the scores of all unproctored tests in another group to compare the participants' performance between proctored and unproctored conditions. According to the independent t test, the scores were significantly higher on unproctored exams in *Adopting the views of other* (t = 2.87, p < .01, η^2=.064). Templer and Lange found no significant difference in all other sections of the test (p was not stated).

To conduct the within-subject analysis, Templer and Lange (2008) utilized a paired t test and found that the students who took the proctored test followed by the unproctored one had significantly higher scores on unsupervised *Creativity based on rules* section (t = -5.68, p < .001, η^2 =.487) and marginally significant higher scores on *Adopting the view of others* (t = -1.98, p = .055, η^2 =.104). The students who took the unproctored test first and the proctored test second had higher scores on proctored *Creativity based on rules* (t = -2.25, p < .05, η^2 =.120) and *Adopting the view of others* (t = -2.48, p < .05, η^2 =.143). Moreover, both control groups also had higher retest scores on the cognitive ability test *Creativity based on rules* (group 1: t = -6.47, p < .001, η^2 =.518, group 4: t = -3.48, p < .01, η^2 =.198). These findings suggested that increase in scores occurred because of repeated testing and not because of exam format effect. No significant differences were found in other sections of the battery test. Templer and Lange concluded that overall they did not find a significant format effect. Moreover, the order in which proctored and unproctored exams were administered did not influence the students' scores.

It is known that the order effect might be especially strong in within-subject studies (Campbell & Stanley1963; Shadish et al., 2002) and should be analyzed in all investigations where the same students are tested more than one time (Shadish et al., 2002; Templer & Lange, 2008). In a study done by Fask et al. (2015), each student took the same exam two times. However, the researcher did not test the order effect, saying that the fact that students did better on the first exam administered in the unproctored format was enough to conclude that cheating took place during the unsupervised exam. This conclusion might be controversial because Fask's et al. students could perform worse on the second exam due to fatigue effect. Templer and Lange (2008) incorporated a different order of proctored and unproctored exams, but their experiment is hardly applicable to a natural educational setting where randomization into test administrations that are part of the curriculum is typically not permissible. Moreover, in natural education settings, all students take all course exams in a certain sequence. The combination of this facts represents another gap, which was addressed in the given within-subject investigation. The curriculum of the department includes two sets of web-based exams. In the first set, the proctored exam is followed by unproctored, in the second set the order is reversed, which allows for studying the order of exam administration. To fulfill the gap, I analyzed whether the order of proctored and unproctored exams administered in a natural educational setting influences exam scores.

Literature Related to what is Known, Controversial, and Needs to be Studied about the Variable Course Delivery Mode

The course delivery mode, face-to-face, hybrid, online, is the second additional independent variable. Beck (2014) analyzed the course delivery mode in three sections of economic course offered in one semester. The researcher utilized a one-way ANOVA and found no significant difference in students' scores across the course delivery modes (*F* =.141, *p* =.869). However, these findings might be controversial because the sample size in the online (*N* =19) and hybrid (*N* = 21) groups was small. Moreover, it is not clear whether a similar effect can be observed at settings different from universities. There was a need to study the course delivery modes with bigger sample size and at different settings. I addressed this need by comparing community college students' scores across the course delivery modes with the size of over 30 students in each mode.

Literature Related to what is Known, Controversial, and Needs to be Studied about the Variable Instructor

The instructor of the course is the last independent variable of the given study. Beck (2014) and Stack (2015) emphasized the importance of controlling for the instructor effect. Both researchers suggested that the easiest way to control for the instructor effect is to analyze the scores of the students taught by one instructor. However, it may result in a small sample size, as it was in Beck's study, and may bring additional bias to the investigation if the instructor is also a researcher. Ladyshewsky (2015) analyzed score of students taught by four different instructors. According to the descriptive statistics, the means of scores were different across the instructors (M_1 =73.9, Mp_2=71.3, M_3 =68.3, M_4 =67.4), but the researcher did not test whether this difference was significant. This gap was addressed in the given investigation. I compared the scores of students taught by seven instructors and investigated whether there was a significant difference between the scores across the instructors. The detailed results of this analysis are given in Chapter 4.

Studies Related to Research Questions

The main goal of the given study is to investigate relationships between the exam format (IV1), proctored versus unproctored, and student scores (DV) when equivalent automatically-scored web-based exams with the same security mechanisms are used. The first research question (RQ1) reflects this goal. Many of the previously discussed studies are related to this question. Harmon and Lambrinos (2008) examined the undergraduate university economics students' scores on the automatically-scored unproctored web-based exam followed by the same web-based exam administered in a proctored format. These one-year-apart exams were secured by randomization of test items and limited time. The unproctored exam was open for 24 hours. Harmon and Lambrinos found that the students who took the unproctored exam (*N* =24) performed better than the students who took unproctored exams (*N* =38) (F_{up} =.02, R^2_{up} =.497, *p* > .05, F_p =35.60, R^2_p =.0008, *p* < .01).

Beck (2014) compared the scores of 80 undergraduate university economics students who took proctored pencil-and-paper multiple-choice exam on campus with 19 students who took an equivalent unproctored web-based exam online. Randomization of test items, restricted time, one question at a time, blocked backtracking, and cheating warning statement posted online were the security mechanisms used by Beck during unproctored exam, which was open for two days. The exams were administered during the same semester. The Beck found no significant difference between the scores (*p* < .05).

Stack (2015) analyzed the undergraduate university criminology students' scores collected during ten semesters. The first five sections (*N* =141) took the proctored pencil-and-paper multiple-choice exam, the last five sections (*N* =170) completed an equivalent unproctored web-based exam. Stack incorporated randomization of test items, restricted time, one question at a time, blocked backtracking, synchronous testing, and lockdown browser as the security mechanisms during unproctored exams. The researcher found no significant difference in scores between the proctored and unproctored groups (R^2 =. 343, b = 1.08, p > .05).

Ladyshewsky (2015) investigated the relationship between the exam format, proctored versus unproctored, and postgraduate university business students' scores collected during nine consecutive terms. The researcher found that the students who took pencil-and-paper multiple-choice test (*N* =136) performed better than the students who completed the unproctored web-based exam (*N* =114) (*Mp* =75.4, *Mup* =72.4). The significance of this difference was not examined. The web-based exam was open for four days. Ladyshewsky utilized randomization of test items, restricted time, one question at a time, blocked backtracking, and higher order thinking questions as the security mechanisms.

The study's second research question (RQ2) is related to the relationship between the order (IV2) in which proctored and unproctored exams are administered and the student scores on the exams. Templer and Lange (2008) investigated this relationship in their laboratory experiment with Singaporean undergraduate university student-volunteers. The researchers compared the participants' performance on the web-based battery test for future employees across four randomly assigned groups. In the first experimental group (*N* = 40) the students took a proctored test followed by unproctored one; in the second experimental group (*N* = 35) the order of the exams was reversed. In the first control group (*N* =38) the students completed both exams in a proctored format; in the second control group (*N* =50), both tests were unproctored. Templer and Lange found a significant difference in the students' scores in some sections of the battery test

($F = 2.94$, $p < .05$, $\eta^2 = .052$), but because the gain in scores in all groups occurred during the retest, the researchers concluded that the order of the test administrations did not influence the performance of the participants. Templer and Lange noted that the practice effect could be the reason of the score increase.

The relationship between the course delivery modes (IV3), face-to-face, hybrid, online, and the student exam scores is reflected in the third research question (RQ3). Beck (2014) studied this relationship in the previously discussed study. The researcher found no significant difference in the students' performance during the unproctored and proctored exams across the course delivery modes ($F = 0.141$, $p = .869$). The last research question (RQ4) is related to the relationship between the instructor of the course (IV4) and the student scores. This relationship was discussed by Ladyshewsky (2015) who compared the score of the students taught by four different instructors. On average, the students of the different instructors performed differently, but Ladyshewsky did not examine whether the difference was significant ($M_1 = 73.9$, $Mp_2 = 71.3$, $M_3 = 68.3$, $M_4 = 67.4$).

The study's research questions have some similarities and differences with the aspects investigated by the researchers discussed above. Similar to the researchers, I examined the relationship between the variables involved in the four research questions, but with the different population of students taught by seven different instructors. The exams involved in the study were web-based and had a combination of security mechanisms different from the combinations used by the discussed researchers. I analyzed the students' scores earned on two sets of proctored and proctored exams administered in reverse order, but taken by each student in the same sequence. The results of the given study were compared with the findings of previous researchers in Chapter 5.

Summary and Conclusion

A major theme emerges from the analysis of the literature discussed in this chapter: With systematically selected security mechanisms, students' scores on unproctored and proctored exams can be similar. It is known that students tend to perform better on unproctored web-based exams with no security mechanisms. The same relationship between the exam format and student scores was observed when only randomization of test items and restricted time were used. It is also known that the researchers who used different combinations of several security mechanisms obtained mixed results.

The studies described above were conducted with undergraduate, graduate, or postgraduate students. Thus, it was not known how the involved variables manifest with the different student population. None of the researchers, who used security mechanisms during unproctored exams, related the choice of the mechanism to the fraud triangle theory considering all its elements, opportunity, need, and rationalization. All of the researchers who utilized a combination of several security mechanisms, used pencil-and-paper mode for proctored and web-based mode for unproctored exams. However, these two modes may not be equivalent. To the best of my knowledge, the order in which proctored and unproctored exams are administered has not been studied with within-subject design in natural educational settings. Most of the researchers analyzed the scores of the students taught by one instructor. It is not known how the presence of several instructors may impact the relationship between the exam format and student scores.

The given study filled the gaps described above. I analyzed the scores of the community college students taught by seven different instructors who used a combination of security mechanisms carefully selected by the department based on the fraud triangle theory. The course delivery mode effect with the entire population of students who took the introductory statistics course offered by the community college was examined. I also investigated whether the order of the proctored and unproctored exams administered in a natural educational setting influences student scores. My findings extended knowledge about credibility of unproctored web-based exams.

The methodology and statistical methods, which were used to analyze all involved variables and fill the gaps in the literature, are discussed in Chapter 3, which follows. This chapter also includes the detailed description of the research design and rationale for its selection, study’s population, the procedure of obtaining archived data, instrumentation, and data analysis. The analysis of threats to validity concludes Chapter 3.

Chapter 3: Research Method

Although academic cheating is a problem in higher education (Bristor & Burke, 2017; Corrigan-Gibbs et al., 2015; McCabe et al., 2012) and can influence exams scores (Arnold, 2016; Corrigan-Gibbs et al., 2015; Fask et al., 2015), cheating was not the topic of the given investigation. The purpose of this quantitative study was to examine whether inconvenient and expensive proctoring is necessary when web-based exams with carefully selected nonbiometric security mechanisms are used. The relationship between the format in which equivalent automatically-scored secured, web-based exams are administered, proctored versus unproctored, and exam scores was examined. If there is no significant difference in students' performance on proctored and unproctored exams, inconvenient and costly proctoring may be avoided. Student performance on proctored and unproctored exams can be influenced by the order in which the exams are administered (Fask et al., 2015), by the course delivery mode (Beck, 2014), and the instructor of the course (Beck, 2014; Ladyshewsky, 2015; Stack, 2015). For this reason, I also analyzed the order, course delivery mode, and instructor effects.

This chapter begins with the explanation of the study's variables, research design and rationale for its selection, which is followed by a detailed description of the study's methodology, target population, and sampling. The description of the procedure of obtaining the archived data and the study's statistical analysis techniques is also in the methodology section. The threats to validity and how they are controlled in the study conclude the chapter.

Research Design and Rationale

The choice of the study's research design is related to the involved variables. The exam format (IV1), proctored versus unproctored, is the main independent variable, while exam score (DV) is the dependent variable. The order in which proctored and unproctored exams are administered (IV2) and course delivery mode (IV3), face-to-face, hybrid, online, are two other independent variables. The instructor of the course (IV4) is the last additional independent variable. To investigate the relationships between these variables, a quasi-experimental one-group sequential design (Thompson & Panacek, 2006) was incorporated.

One-group sequential designs consist of one group of participants involved in all treatments administered in a certain sequence (Thompson & Panacek, 2006). In the given study, the students, whose scores were analyzed, were considered as one group of individuals who participated in all conditions by taking two sets of web-based exams in a certain sequence. In the first set, the proctored exam was followed by unproctored; in the second set, the order of the exams was reversed. In this design, each student was perceived as his/her own control. This within-subject approach allows for answering the first two research questions, described as follows. To test RQ1, whether there is the relationship between the exam format, proctored versus unproctored, and the exam scores, the differences between the individual scores on proctored and unproctored exams was analyzed. To assess RQ2, whether there is the relationship between the order in which the proctored and unproctored exams are administered and exam scores, the differences in the individual scores when the proctored exam was administered first were compared with the differences in the individual scores when the proctored exam was administered second. The department archived scores of the proctored and unproctored web-based exams administered in face-to-face, hybrid, and online sections of the course taught by seven different instructors. This information allows for between-subject comparison in assessing the course delivery mode (RQ3) and the instructor effects (RQ4).

The choice of the quasi-experimental design is related to constraints of a natural educational setting. The scores archived by the department were collected at as a part of a regular educational practice where random assignment to control condition, proctored, and experimental condition, unproctored, was not possible. Moreover, the students, whose archived scores were analyzed, were self-enrolled in their classes, which did not allow for random assignments with respect to either the course delivery mode or instructor. There were also time constraints. To obtain sufficient sample size, the scores archived during several semesters were requested. The department has been administering proctored and unproctored web-based exams in all course delivery modes since 2015. However, in Fall 2017, the college switched from Moodle LMS to Canvas LMS, and the web-based exams' structure was changed due to differences between the LMSs. To avoid possible new LMS effects, only the scores archived from Fall 2015 to Summer 2017 were requested.

The choice of the study's design is consistent with research designs needed to advance knowledge about web-based testing in higher education. Quasi-experiments with controlled validity threats are the best approaches in studying cause-effect relationships in natural educational settings when randomization is not possible or unethical (Kim & Steiner, 2016; Shadish, 2011). Validity threats of quasi-experiments can be effectively reduced, eliminated, or assessed by using a combination of within- and between- subject design elements (Charness et al., 2012). The given study's design incorporates both within-and between-subject elements. The within-subject factor of the design, one group of the participants in which each student serves as his or her own control, eliminates selection biases (Thompson & Panacek, 2006). The between-subject design elements, exam scores of the students enrolled in online, hybrid, and face-to-face sections taught by seven instructors, allow for assessing possible validity threats related to course delivery mode and instructor effects.

A quasi-experimental one-group sequential design was used by Fask et al. (2015) in a natural educational setting in their analysis of scores on unproctored and proctored exams in an introductory statistics university course. However, although the researchers mentioned the importance of assessing the order effect (Fask et al. 2015), the design of their study included only one set of exams, in which unproctored exam followed by proctored. In the given study, the second set of exams has the reverse order, which allows for examining the order effect. To the best of my knowledge, this study is the first study that examined the order in which proctored and unproctored exams are administered in a natural educational setting. Beck (2014) studied the course delivery mode effect, but the sample sizes of 19 in the online section and 21 in the hybrid sections were small. Ladyshewsky's (2015) design included scores of students taught by three different instructors, but he did not use any inferential statistical analysis to test the instructor effect. In the given study, I tested the course delivery mode effect with larger sample size of 664 face-to-face students, 91 hybrid students, and 55 online students. Additionally, I tested the instructor effect statistically by comparing the students' performance across the seven different instructors. Thus, the quasi-experimental one-group sequential design has all elements needed to address the gaps in the studies of previous researchers and advance knowledge about web-based exam administrations in natural educational settings.

Two pairs of secured, web-based exams administered in proctored and unproctored formats in different order can be perceived as an intervention. The department where the study took place decided to implement these exams in all web-based sections of the introductory statistics course. This investigation was designed to inform whether the security mechanisms selected by the department can substitute for proctoring.

Methodology

Quantitative strategy of inquiry is the methodology of the given investigation. This choice of methodology fits the main goal of the study to examine a relationship between web-based exam format, proctored versus unproctored, and exam scores collected in numerical form. This section of the chapter begins with a detailed description of the study's population and sampling, followed by the explanation of the data collection procedure and instrumentation. The description of the data analysis plan concludes the section.

The Study's Population

The study's setting is a Californian suburban community college, which serves 9,000 students every semester. In Fall 2015, the student population was identified as 29 % White, 19% Hispanic, 18% Asian, 18% Multiracial, 8% Filipino, 3% African American, 3% Pacific Islander, and 2% others. The average age of the students was 26; 36 % of all attendees were full time students; about 52% were males and 48% were females. The college offers transfer programs in 11 subject areas, nine of which have Introductory Statistics as a transfer requirement. Art and Foreign Languages and Pure Mathematics, Engineering, and Computer Science constitute the remaining two subject-areas.

Introductory Statistics taught in the department is a typical community college introductory level statistics course that satisfies all prerequisites required by universities for transfer. Twenty-one out of 23 introductory statistics sections offered by the college every year are web-based sections in which the course content is delivered through the college management system in face-to-face, hybrid, and fully online modes. The remaining two sections are not web-assisted and offered in a classroom without computers. There are 10 web-based face-to-face sections and one hybrid section every fall and spring semesters, and one fully-online section every summer. All individuals enrolled in the online, hybrid, and web-assisted, face-to-face sections from Fall 2015 to Summer 2017 were the target population of the study.

Sampling and Sampling Procedures

The census sample, the collection of the entire population under study, was utilized in the given investigation because the web-based exams' scores of each student in the identified population was analyzed. This sampling strategy was determined by the goals of the study, the type of data archived by the department, the setting where the study took place, and the desired statistical power of the test. To answer the research questions of the investigation, the numerical scores of proctored and unproctored web-based exams administered in a different order in online, hybrid, and face-to-face introductory statistics sections taught by different instructors were needed. The department, which has been archiving all exams' grades in all web-assisted sections for educational purposes since Fall 2011, could provide the needed numerical scores. The instructors of the department decided to implement two sets of proctored and unproctored exams in all web-assisted courses in Fall 2015. In the first set, the proctored exam was followed by the unproctored; in the second set the order was reversed. The scores obtained by each student on these exams beginning Fall 2015 was the part of the data archived by the department. These scores and some student demographics such as GPA, age, and major were requested for the data analysis.

The students, whose scores were analyzed, were not randomly enrolled in their classes: the enrollment procedure established by the college is based on self-selection. Thus, the participants were not randomly assigned to the course delivery mode or instructor. For this reason, a true experimental design with random assignments into groups, a frequently preferable approach due to its higher internal validity, was not possible for this investigation. However, the findings of carefully designed nonrandom studies can be generalized to similar courses offered in similar institutions (Shadish, 2011).

Introductory Statistics is the only course in the department that has been offered in web-assisted face-to-face, hybrid, and online formats on a regular basis during the last 6 years. All sections of the course are scheduled in a computer classroom that allows for the use of technology needed for proctored web-based exams. Moreover, the instructors who teach these sections use the same curriculum by utilizing the same calendar, materials, and assessments, which eliminate the curriculum effects. For this reason, the study was delimited to all students who took web-assisted Introductory Statistics from Fall 2015 to Summer 2017. The scores of all these students were analyzed.

At the institution where the study took place, students can withdraw from class by specific deadline. Because the second set of the web-based exams took place after the drop deadline, the number of students who completed the exams in both sets was smaller than the number of students who took the exams in the first set. However, a sufficient population size is needed to achieve the desirable statistical power of the appropriate statistical analysis techniques (Cohen, 1988; Murphy, Myors, & Wolach, 2014). For this reason, the scores of the students who took the first set and then dropped the class were kept in the analysis. To test the format effect (IV1), a repeated measures ANOVA was used to analyze the Set 1 scores first and then the Set 2 scores. After that, a mixed ANOVA was applied to the scores of the students who took both sets to test the order (IV2), course delivery mode (IV3), and instructor (IV4) effects. The scores of four exams

(Set 1 proctored, Set 1 unproctored, Set 2 unproctored, Set 2 proctored), three levels of course delivery mode (face-to-face, hybrid, online), and seven different instructors were involved in the study. Because the scores of the entire population of web-based introductory statistics students were analyzed, no power analysis to determine sample size was needed.

Intervention

Although the given study utilized archival data, the departmental implementation of web-based exams with systematically selected security mechanisms can be perceived as an intervention. The proctored format of the exams is the control condition, while the unproctored format is the experimental one. Each student went through both conditions by taking two proctored and two unproctored exams in the following sequence: Set 1 proctored, Set 1 unproctored, Set 2 unproctored, Set 2 proctored. The department 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. 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 the second set was administered in an unproctored environment, while the alternative form of this exam was proctored. This choice of the format of each exam, occurring in a natural educational setting, can be considered as a natural experimental manipulation of the variable order. The detailed description of how the intervention in a form of secured proctored and unproctored web-based exams was designed, implemented, and administered by the department is provided below.

In 2015, the third version of Moodle with enhanced testing features was released (Moodle, 2015). The department realized numerous advantages of the improved web-based testing available in the college LMS Moodle and decided to substitute some regular pencil-and-paper tests by proctored and unproctored web-based exams in all Introductory Statistics sections offered in web-assisted face-to-face, hybrid, and online modes. The choice of the content of the exams, their number, order, and the time interval between the tests depended on the departmental curriculum, procedures, and policies. Some decisions made by the instructors were informed by research on cheating reduction mechanisms related to web-based testing and strategies that reduce retesting effects. The introductory statistics faculty developed four web-based exams and several web-based practice tests in the college LMS and incorporated carefully selected security mechanisms based on best practices and research. The structure, duration, and the security mechanisms of all eight exams were identical. The purpose of the practice tests was to familiarize students with the structure, duration, and question type of exams and minimize rationalization and need to cheat. Synchronous test administration, restricted time, blocked backtracking and feedback, randomization, one question per page, clearly defined academic integrity policies, and cheating warning statements are security mechanisms utilized by the department. A detailed description of these security mechanisms, their purposes, rationale for their selection, and relation to opportunity, rationalization, and need cheating factors are provided in Chapter 2 and the corresponding tables in Appendix A and Appendix B.

The first two web-based exams developed by the instructors covered inferential procedures for population proportions; the second pair of exams was about inferential procedures with population means. All four exams had 23 items: three drop-down, four multiple-choice, and 16 short answer questions. The detailed description of how the exam items were created is provided in the Instrumentation section of this chapter. The samples of exams' questions are located in Appendix C, Appendix D, and Appendix E.

The first web-based exam was administered in class at the middle of the semester and was proctored by the instructors. The faculty decided that by the middle of the semester the students are well-familiar with the LMS and the procedure of web-based exam administration, which they learn by taking the practice web-based exams. The department has a policy of cumulative exams such that several items on each exam cover concepts similar to the concepts covered on previous tests. The faculty divided a pencil-and-paper midterm, which was used before the web-based testing incorporation, into two sections: all cumulative questions, like the questions of the first web-based exam, were moved to the second web-based test, which is administered outside of class in an unproctored environment. All questions on the new concepts stayed in the smaller pencil-and-paper part. Thus, the second web-based exam is an alternative version of the first one and administered in 7-10 days after the initial test. The instructors selected the interval of 7-10 days based on Spitzer's (1936) findings about forgetting of a large portion of initial reading in 7 days. The students did not know that the second exam was an alternative to the first one, and no individual feedback was given between the exams. All these aspects reduce practice and fatigue effects (Catron & Thomson, 1979; Spitzer, 1936; Villado et al., 2016).

The second pair of web-based exams was administered at the end of the semester in 1 month after the students took Set 1 exams. Although this test-retest interval was dictated by the course curriculum, Falleti et al. (2006) found no significant practice effect on web-based exams with test-retest interval of 1 month. The second pair of tests had six questions identical to the Set 1 exams. The other 17 questions covered new but equivalent concepts at the same level of difficulty and had identical structure. The alternative version of the unproctored exam in Set 2 was administered as a proctored web-based portion of the final exam in 7-10 days after the Set 2 initial test. Thus, the exams in Set 2 were administered in the reverse order, which allowed for examining the order effect. The detailed diagram of the intervention is provided in Figure 1 in Chapter 2.

Data Collection

The archival scores collected by the department during regular web-based exams administered in a natural educational setting were analyzed in the study. For this reason, there was no actual recruitment of the participants, and no consent forms were necessary. The students self-enrolled into Introductory Statistics based on their academic goals, completed the first day survey administered through the LMS, worked on the assignments, and took all course assessments, including the four web-based exams. After each semester is over, the statistics coordinator of the department downloads all web-based exams' scores and the survey data automatically recorded by the LMS, adds college GPA provided by the institution, and collects the pencil-and-paper tests' scores and course grades from faculty. The first-day survey includes questions on course delivery mode, instructor, age, gender, major, academic units, working hours, college grade level, time of the day when the section is offered, and other personal student characteristics. The department utilizes the survey data to create problems and projects; the exam scores and course grades are used for student learning outcomes analyses and reports.

When my Walden IRB application was approved, to answer the study's research questions and test the hypotheses, I requested the scores obtained by the students on all web-based exams involved in the study and instructor and course delivery mode information. To describe the demographics of the study's population, I also requested student GPA, age, gender, major, college grade level, academic units, and the number of working hours. Several researchers found that GPA and age can influence cheating behaviors (Beck, 2014; Harmon & Lambrinos, 2008; Gallant et al., 2015; Ladyshewsky, 2015; Stack, 2015). For this reason, to interpret the study's results, GPA and age were utilized to identify whether the students were different across the groups with respect to the course delivery mode and instructor. To address the attrition bias, I asked for the dropout rate of statistics students in the 2014-2015 academic year. The Introductory Statistics coordinator of the math department provided the requested data in a spreadsheet with all identifying information removed to protect the identities of the students and instructors. The names of the instructors were also removed and a code entered.

To get the permissions to utilize archival data and gain access to them, I submitted a summary of my proposal with a detailed description of the study's design and data collection to the Institutional Review Board at the college where the study took place. The IRB committee permitted me to conduct the study and use the archival data described above for the analysis and reports.

Instrumentation

A team of introductory statistics instructors of the department, experts in the discipline, developed the Introductory Statistics inquiry-based curriculum with the corresponding lecture notes, homework assignments, project activities, and pencil-and-paper exams. Each semester, during the first week of classes, the department surveys all statistics students on age, major, instructor, semester, course delivery mode, the number of academic units, the number of hours spent working, political views, the number of languages spoken, and other personal characteristics. When the college adopted the LMS Moodle in 2008-2011, the survey and course materials, including homework, were posted on the course websites. Currently, the web-based survey instrument administered through the LMS has 30 drop-down and checkbox questions on demographic and student personal characteristics described above. The statistics coordinator downloads the data, adds college GPA provided by the institution, and saves it on an Excel spreadsheet to be used for educational purposes and program evaluations. The instructor and course delivery mode information was requested to sort the student exam scores with respect to these categories and test the course delivery mode (IV3) and instructor (IV4) effects. Student GPA, age, major, semester, college grade level, the number of working hours a week, and number of academic units were requested to describe the study's population. GPA and age were utilized to compare the participants across the groups. However, the main study’s instruments are web-based exams, which were created by the department in accordance with the best practices described below.

The Best Practices of Exam Development

Van der Vleuten (1996), the author of an excellent assessment framework, identified feasibility, cost effectiveness, acceptance, educational impact, and reliability and validity of an exam as main criteria for high-quality assessment. Feasibility and cost effectiveness corresponds to practicality and suitability of a test to the given setting, while acceptance demonstrates whether all participated parties accept the process of test administration and its results. Educational impact indicates that assessment motivates and engages students to learn. Reliability describes the precision of the test and its reproducibility, while validity reflects whether the exam accurately measures what it is supposed to measure (Van Der Vleuten, 1996).

While Van Der Vleuten (1996) focused mostly on general factors of effective assessment, Haladyna, Downing, and Rodriguez (2002) developed a taxonomy of 31 writing guidelines for effective multiple-choice questions based on an analysis of 27 textbooks and 27 research articles published after 1990. Each multiple-choice item has a question or stem and list of possible responses, out which one is correct and others are distractors. Item specific important content, independence of items from each other, simple vocabulary and wording, formatting of responses vertically, correct grammar, spelling, and punctuation, and minimization of needed reading are some of guidelines related to content, formatting, and style. In writing the stems and response choices, Haladyna et al. (2002) recommended including the central idea in the stem, not responses, and formulate the stem and choices positively. The response choices should have the same grammatical structure and lengths; the distractors have to be plausible. The phrases “none-of-the-above” and “all-of-the-above” should be avoided. Stems and responses should not be overly wordy and contain only relevant not repetitive information and be free of logical clues such that the instrument measures student knowledge of the subject matter and not reading comprehension or test-wiseness (Haladyna et al., 2002).

Poorly created exams with incorrect grammar, not plausible responses, and more than one correct answer can penalize knowledgeable students, while weak students can unfairly benefit from logical cues (Dell & Wantuch, 2017; Downing, 2005; Tows, 2014). Additionally, inadequately written exam questions of any type may assess reading comprehension rather than content (Dell & Wantuch, 2017). These flaws can lead to inability to discriminate between high and low performing students (Dell & Wantuch, 2017; Downing, 2005; Towns, 2014). Thus, item flaws may reduce validity and reliability of exams. Several researchers have studied exam item flaws and their impact (Cassels & Johnstone, 1984; Downing 2005; Odegard & Koen, 2007; Rodriguez, 2005).

Research Related to Exam Items’ Flaws

Cassels and Johnstone (1984) investigated the influence of wording of exam questions on 3,600 chemistry students’ performance. The researchers created two equivalent tests but with several differently worded questions and administered them to two groups of students during three consecutive years. Some of the questions, control questions, were identical on both exams. The researchers compared the control questions’ scores between the groups and found no significant difference (*p* and other statistics were not provided); this result was observed during each year. Because the scores on the controlled questions did not differ significantly, Cassels and Johnstone assumed that any differences between pairs of matched questions should be due to wording and not due to differences among the groups.

Cassels and Johnstone (1984) found that the students performed significantly better with the use of simpler words: the proportion of students who answered correctly the revised questions increased from 49% to 56% ($p < .05$). When the words “least concentrated” were changed to “most abundant,” the significant improvement in correct responses from 59% to 75% ($p < .05$) was observed. The researchers suggested that less thinking might be needed to answer the questions with the phrase “most abundant.” The students’ performance significantly increased from 24% to 80% ($p < .05$) when all negative forms were changed to positive. The change was especially noticeable in questions with double negatives. The wordy questions were harder to answer than the questions with the same complex phrase divided into a few shorter sentences (47% vs. 67%, $p < .05$). Cassels and Johnstone explained that wording impacts thinking: simpler phrases, short sentences, and nonnegative forms requires less thinking stages, which may improve performance. Cassels & Johnstone concluded that to assess students’ knowledge accurately, the instructors should use exam language free of complex sentences and negative forms.

Unlike Cassels and Johnstone (1984) who studied the influence of English words and phrases on student performance, Downing (2005) investigated the effect of writing flaws on item difficulty, test reliability, and discrimination. Four science exams from different science discipline were randomly selected for the study; the examinees, medical students, were the same for some exams (Downing, 2005). Three independent experts, blinded to student performance data, used Haladyna's et al. (2002) taxonomy of 31 criteria to classify the exams' items as *standard,* without any flaws, and *flawed*, if at least one criterion was violated. Unclear stated questions in a stem, "none-of-the-above" response option, and negative-stem item were some of the flaws. There were 100 flawed items out of 219 total questions (46%). All involved exams' items were multiple-choice questions with five response options and one correct answer. All exams were proctored pencil-and-paper tests with scantrons and had restricted time (Downing, 2005).

Downing (2005) found that internal-consistency reliability of scores was between .66 and .78 and mean point-biserial item discrimination indices were between .18 and .21. The mean item difficulty ranged from 4% to 15% points more difficult for flawed items than for standard ones; the passing rate from 2% to 5% point lower for the flawed items. The standard and flawed score correlation ranged from .45 to .68 ($p < .01$). About 14% of the students who completed all test items answered the standard questions correctly but answered the flawed ones incorrectly. The effect of flawed items on reliability was mixed: on three exams the reliability score was larger for flawed items and only on one test flawed items were less reliable, which could occur due to the systematic error of the measurement. Downing concluded that the flawed test items influenced item difficulty, discrimination, and student passing rate.

Some researchers specifically studied flaws related to response options of multiple-choice questions. Rodriguez (2005) conducted a meta-analysis on the most effective number of responses in multiple choice questions. The researcher analyzed 27 studies published between 1925 and 1999 with 56 trials total and 1657 participants. All these studies employed either random assignments into groups or pre-posttest designs with one group of students. The researchers conducted item difficulty, item discrimination, test score reliability analyses, compared the effects, and estimated covariance after four methods in changing of number distractors were utilized: random removal (N =26), removal of ineffective distractors (N =19), removal of the most effective distractors (N =3), and adding distractors (N =1). There was no relation between item difficulty and item discrimination with respect to the deletion method. However, after the researcher deleted most ineffective and most attractive distractors, for example, from 4 to 3 options, the mean item difficulty changed significantly making the question easier (p =.01) and a decrease in item discrimination was detected (p < .05). Reduction of a number of responses decreased reliability with the largest decrease in the reduction of response options from 5 to 2 (p < .05). The researcher concluded that three response options, single answer and two distractors, is sufficient; however, four or five options can increase content coverage and improve test score reliability. Rodriguez suggested that less than three responses in multiple choice questions may increase guessing and decrease content coverage, while more than five responses can increase item difficulty and number of logical clues. Thus, based on the findings of Dell & Wantuch (2017), Rodriguez

(2005), and Towns (2014), it can be concluded that the number of responses between three and five is the most optimal.

Odegard and Koen (2007) investigate a relation between “none of the above” as a correct and incorrect response and student performance. Thirty-two undergrad students were asked to read 18 out of 36 nonfiction passages and take a 72-item multiple-choice test. In this test, 36 questions were related to the read passages and 36 questions to the unread passages. The passages were counterbalanced: each excerpt served as read and unread the same number of times. Odegard and Koen created the test items which had 4-response options and 5-response options with the fifth option of “none-of- the-above.” In half of the questions with five responses, the “none-of- the-above” option was the correct choice and in another half incorrect. In Odegard and Koen’s study, the retest, which had some previously tested questions (control) and some previously untested questions (critical), was administered in five minutes after the initial test.

Odegard and Koen (2007) utilized a mixed 2 (passage status: read, unread) x 2 (item type: critical, control) x 2 (question type: 4 responses, 5 responses with incorrect “none-of-the-above” choice) ANOVA and, as it was expected, found that the students scored higher on the questions for read excerpts (*M* =.36, *SD* =.18) than on questions from unread excerpts (*M* =.14, *SD* =.09, *F* (1,30)=79.28, *MSE* = .019).There was no effect with respect to the question type or any interaction: a testing effect was present on the retest for all questions, including an incorrect “none-of-the -above” responses (*F*s <1.28). Similarly, the researcher ran a mixed 2 (passage status: read, unread) x 2 (item type: critical, control) x 2 (question type: 4 responses, 5 responses with correct “none-of-the-above” choice). There was a significant question type by item type interaction (*F* (1, 30) =36.52, *MSE*=.006). The students in the control group without “none-of-the-above” responses perform significantly better on the items that were assessed before (*M* =.33, *SD* =.17) than on the items that were not assessed before (*M* =.20, *SD* =.12, *F* (1, 30) =50.61, *MSE*=.006). The students in the experimental group with the correct “none-of-the-above” responses did not perform better on the items that were assessed before (*M* =.15, *SD* =.10) than on the new items (*M* =.18, *SD* =.10, *F* (1, 30) =2.05, *MSE* =.006, p > .05). Thus, there was no testing effect in the correct “none-of-the-above” responses group. The researchers also analyzed the participants’ performance on the initial test and found that for the read passages the scores were lower when the incorrect “none-of-the-above” responses were included (*M* =.55, *SD* =.16) than when they were not used (*M* =.69, *SD* =.15, *F* (1, 30) =15.80, *MSE* =.01). The corresponding difference was not significant for unread passages (*F* (1, 30) =1.00, *MSE* =.01, p > .05). Odegard and Koen concluded that the inclusion of “none-of-the-above” responses influenced student performance and did not recommend using them in multiple-choice questions.

In addition to the structure of responses of multiple-choice items, adequate content of the questions is needed. Towns (2014), Dell and Wantuch (2017), Kibble (2017) reviewed the best practices of the development of high-quality exam questions accumulated by previous research and suggested adding to the list a procedure, in accordance to which assessments can be aligned with course objectives. The researchers emphasized that the main goal of summative tests is to assess knowledge of examinees through well-design test items. For this reason, any assessment should be aligned with the course objectives, and this alignment should take place with respect to the course content and cognitive levels (Dell & Wantuch, 2017). To achieve this alignment, Towns (2014), Dell and Wantuch (2017), and Kibble (2017) recommended using taxonomies of educational objectives.

Taxonomies of Educational Objectives

Numerous taxonomies of educational objectives have been developed to accommodate the needs of specific academic subjects and fields (Darwazeh & Branch, 2015; Dell & Wantuch, 2017; Haladyna et al., 2002). Bloom's (1964) taxonomy and the revised Bloom's taxonomy (Anderson, Krathwohl, & Bloom, 2001) are two taxonomies of educational objectives commonly used by educators (Darwazeh & Branch, 2015; Dell & Wantuch, 2017; Towns, 2014). Bloom's (1964) taxonomy has six levels associated with cognitive characteristics of learning, the complexity of which increases with each level: *knowledge*, *comprehension*, *application*, *analysis*, *synthesis*, and *evaluation*. Knowledge involves remembering of facts, terms, categories, rules, and theories, while comprehension is related to understanding of the remembered facts and rules through organizing, comparing, and interpreting. Application constitutes of solving problems using acquired knowledge, and analysis deals with breaking information into parts, examining relationship between them, and making appropriate inferences and generalizations. Bloom's synthesis refers to creating a structure or combining parts into one whole, while evaluation constitutes of opinions and judgments of ideas or work.

However, the categories of Bloom's taxonomy describe only the knowledge dimension and omit factors related to the cognitive process dimension (Krathwohl, 2002).

Anderson et al. (2001) realized this weakness of the original Bloom's taxonomy and rearranged the elements of the framework. First, Anderson et al. included the procedural dimension described in verbs and changed synthesis to creation. The revised taxonomy constituted of *remember*, *understand*, *apply*, *analyze*, *evaluate*, and *create* categories. The category create was related to generating new ideas, constructing new plans, and producing new products. Second, the researchers added knowledge dimension described in nouns: *factual knowledge*, *conceptual knowledge*, *procedural knowledge*, and *meta-cognitive knowledge*. Knowledge of terminology belongs to factual knowledge; classifications, categories, principles, generalizations, theories, and models are aspects of conceptual knowledge. According to Anderson et al., knowledge of algorithms, techniques, methods, and processes belong to procedural knowledge, while metacognitive knowledge involves knowledge of cognitive and strategic tasks.

Darwazeh and Branch (2015) noted that the revised Bloom's taxonomy does not reflect recent findings about human information and cognition processes and suggested a new revision of it. According to Darwazeh and Branch, meta-cognition is not a type of knowledge, but a process that represents thinking. The authors also realized that the *organizing* process, which Bloom (1964) called synthesis, is entirely missing in Anderson's et al. (2001) taxonomy. The organizing, which usually takes place after analysis, and creation used by Anderson et al. (2001) are not equivalent because creation is more complicated than organizing (Darwazeh & Branch, 2015). Moreover, according to the component display theory developed in 1983, remembrance has two levels: (a) remember specific information such as names, symbols, terms and (b) remember general information such as concepts, principles, and procedures. Darwazeh and Branch (2015) also claimed that application process occurs after analysis and organizing, and knowledge of principles should be added to the knowledge dimension. Thus, Darwazeh and Branch's revision of the revised Bloom's taxonomy included nine cognitive processes, the difficulty of which increases with each component: *facts' remembrance*, *generalities' remembrance*, *comprehension*, *analyzing*, *organizing*, *application*, *evaluation*, *creation*, and *meta-cognition*. The knowledge dimension consisted of four parts, also in increasing order of difficulty: factual knowledge, conceptual knowledge, principles knowledge, and procedural knowledge. Darwazeh & Branch suggested that their revision can be more effective in the development of curriculum and assessment, especially in subjects, learning and understanding of which require complex cognitive processes.

Many objectives of the introductory statistics course involved in the given investigation are focused on statistical reasoning and interpretations that require higher order thinking and complex cognitive processes. To align these objectives with the exams, some tests' items should focus on assessing critical thinking. However, it is widely known that the development of higher order thinking items might be challenging, especially with automatically-scored questions (Tractenberg, Gushta, Mulroney, & Weissinger, 2013). Whether and how exam items can be designed to test higher order thinking was studied by several researchers (Dell & Wantuch, 2017; Jensen et al., 2014; Nash & Krauss, 2015; Tractenberg et al., 2013).

Research Related to Designing Higher Order Thinking Questions

Tractenberg et al. (2013) investigated whether cognitive complexity in taxonomies of educational objectives can be separated from item difficulties. The researchers also aimed to develop a framework for training instructors in writing higher order thinking exam items. Tractenberg et al. analyzed 252 multiple-choice questions from three web-based exams administered in a graduate psychology course during the same semester. All exam items had five response options, were scored dichotomously, correct or incorrect, and randomized. The researchers developed a cognitive complexity matrix that corresponded to the cognitive levels of Blooms' (1964) and Anderson et al. (2001) taxonomies: each exam item was classified using the revised Bloom's taxonomy, and the corresponding category from the original Bloom's taxonomy was assigned as well. Tractenberg et al. specified that three independent raters, experts in assessments and the subject matter, rated each exam item by applying the developed cognitive complexity matrix.

Tractenberg et al. (2013) found that from 41% to 48% of all questions were rated as items with low difficulty at remember level; from 2% to 3% were rated at analyze level, and the remaining questions were equally distributed between understand and apply levels. No items were classified as evaluate and create. The researchers applied a multiple regression model and found that cognitive complexity at remember, understand, apply, and analyze levels accounted for little variation in item difficulty ($R^2{}_{E1}$ =.027, $R^2{}_{E2}$ =.036, $R^2{}_{E3}$=.014). Tractenberg et al. concluded that there was no relationship between difficulty of the item and cognitive complexity. Therefore, cognitive complexity and item difficulty are independent and can be modified separately, which implies that the cognitive complexity of exam items can be increased without making questions more difficult. Tractenberg et al. also mentioned that faculty can be trained to write higher order thinking exam items by using the cognitive complexity matrix.

Unlike Tractenberg et al. (2013), whose exam included only multiple-choice items, Jensen et al. (2014), in their web-based assessments incorporated not only multiple-choice items, but also fill-in-the-blank, and short-answer questions. Jensen et al. developed two sets of quizzes and tests to examine whether the use of higher cognitive levels of exam items can result in a deeper comprehension of the subject matter. The first set constituted from weekly quizzes and three unit exams, all questions of which were designed at the remember level of the revised Bloom's taxonomy. The second set included weekly quizzes and three exams on the same concepts but at the apply, analyze, and evaluate levels. The purpose of the assessment, types of questions used, and limited time were not suitable for utilizing items at the create category. The researchers also did not include understand questions to clearly distinguish between the lower levels (LL) of thinking items and higher level (HL) of thinking items. The LL items tested how well the students memorized terms and definitions related to a concept, while the HL questions required critical thinking in utilizing the terms and definition of the same concept in analysis, application, and evaluation. Three independent experts trained in Bloom's taxonomy classified the items as LL or HL and made sure that the both sets of assessments cover the same concepts and have the same level of difficulty.

Jensen et al. (2014) used a quasi-experimental nonequivalent groups design. The first set of assessments was administered to a biology section of 84 undergraduate university students; the second set to another biology section of 85 students. Both sections were taught by the same instructor who used the same curriculum, including inquiry-based assignments, and completed the same final exam. The final exam, which tested academic achievement of the participants, had 20 LL items and 21 HL items, which were similar to the questions on the quizzes and unit exams, but had different themes and numerical values. All assessments were proctored and administered outside of the class in the university proctoring center. To measure the initial reasoning ability of the students, Lawson's Classroom Test of Scientific Reasoning Skills was administered to both classes at the beginning of the semester. Jensen et al. recorded the reasoning score of each student.

Jensen et al. (2014) found that LL group had slightly lower reasoning score than HL group (M_{LL} =17.8, M_{HL} =19.4, t (167) = 2.61, p =.01, η_p^2 =.039). However, both groups were capable of learning science concepts successfully because the reasoning score of at least 14 indicates well-developed reasoning skills (Jensen et al., 2014). According to a 3 (exam number) x 2 (experimental conditions: LL, HL) mixed ANCOVA results, with the reasoning score as a covariate, there was no significant difference in mean scores across the LL and HL groups on all three exams (F (2,332) =1.44, p > .23, η_p^2 =.009). To measure academic achievement, the researchers applied 2 x 2 mixed ANCOVA with the final exam questions' level as a within-subject factor, type of assessment through the semester as a between-subject factor, and the science reasoning score as a covariate. The main effect, type of assessment, was significant (*F* (1,166) = 7.15, *p* =.008, η_p^2 =.041). The students who took HL unit tests performed better on the final exam than the students who took LL unit tests. Moreover, the HL group scored significantly higher on both LL (*F* (1,166) = 4.19, *p* <.05, η_p^2 =.025) and HL (*F* (1,166) = 6.32, *p* <.02, η_p^2 =.37) final exam questions than the LL group. Jensen et al. found that the science reasoning score was positively related to the final exam scores (*F* (1,166) = 41.50, MSE=.02, *p* <.0001, η_p^2 =.200).

Jensen et al. (2014) claimed that the use of HL test items increased student involvement into the learning process. The researchers suggested that the students in HL group adjusted their study strategies to match the higher cognitive level of weekly quizzes and unit exams. Moreover, Jensen et al. noted that their findings paralleled the hierarchical assumptions of Bloom's taxonomy about knowledge processes: preparation for and engagement with HL exam items automatically foster mastering of the LL concepts. The researchers concluded that it is possible to align multiple-choice, fill-in-the-blank, and short-answer exam questions with inquiry-based course objectives, using higher levels of Bloom's taxonomy. This alignment improves the quality of assessment and provides opportunities for students to demonstrate what they learned. Jensen et al. concluded that assessments designed at higher levels of Bloom's taxonomy direct students' learning, engage them into more active studying, and potentially lead to deeper understanding of the subject matter.

Unlike Jensen et al. (2014), who did not provide a detailed description of exam questions' development, Nash and Krauss (2015) clearly explained how they aligned assessment with cognitive dimensions and learning objectives in developing a module exam for undergraduate course in Information System. Nash and Krauss given a step-by-step method of designing exam items based on the revised Bloom's taxonomy. The researchers numbered the cognitive process dimension of the revised Bloom's taxonomy as 1. Remember, 2. Understand, 3. Apply, 4. Analyze, 5. Evaluate, and 6. Create. Similarly, the knowledge dimensions were named A. Factual knowledge, B. Conceptual knowledge, C. Procedural knowledge, and D. Meta-cognitive knowledge. Then they mapped each exam question against the learning objective of the module, the corresponding knowledge dimensions and cognitive processes required to answer the question, and the number of points being awarded. For example, if the corresponding learning objective required procedural knowledge to apply a procedure described in the question, the researchers classified the exam item as C3. The researchers assigned two points for questions on which students could obtain a partial credit and one point on questions without partial credit. Nash and Krauss suggested that exam items carefully designed based on taxonomies of educational objectives and aligned with the course learning outcomes could measure student knowledge more reliably and validly.

Dell and Wantuch (2017) emphasized that any development of exam questions begins with the list of learning outcomes the exam is covering. Then each test item is aligned with the corresponding outcome using taxonomies of educational objectives. To simplify the process of generating and writing exam items, the researchers recommended breaking down major cognitive processes of taxonomies into three major categories: knowledge-recall, interpretation-application, and problem solving-evaluation. According to Dell and Wantuch, further classification and tailoring of taxonomies might be needed to match the specific course objectives.

The previous theorists, researchers, and educators identified that effective assessments are feasible, cost effective, accepted by all stakeholders, reliable and valid, and have an educational impact (Van der Vleuten, 1996). High-quality exam items have correct grammar, punctuation, and spelling (Dell & Wantuch, 2017; Cassels & Johnstone, 1984; Haladyna's et al., 2002; Towns, 2014), include only necessary information, are not repetitive, and free of logical cues (Dell & Wantuch, 2017; Towns, 2014). Moreover, the exams' questions are positively stated and independent of each other (Dell & Wantuch, 2017; Cassels & Johnstone, 1984; Towns, 2014). The response options, including distractors, are plausible (Dell & Wantuch, 2017; Haladyna's et al., 2002), stated in parallel format by using the same verb tense and have equal length (Dell & Wantuch, 2017; Haladyna's et al., 2002), contain three to five answer responses (Dell & Wantuch, 2017; Rodriguez, 2005), do not include "none-of-the-above" (Haladyna's et al., 2002; Odegard & Koen, 2007) or "all of the above" (Xu, Kauer, & Tupy, 2016) options, and have only one correct answer (Dell & Wantuch, 2017; Haladyna's et al., 2002). To prevent fatigue effects, the optimal number of items on a high-quality test is about 25 (Haladyna et al., 2002) with at least three minutes per higher level of thinking question (Ladyshewsky, 2015). Each exam item is aligned with course learning outcomes (Dell & Wantuch, 2017; Kibble, 2017; Nash & Krauss, 2015) through the use of an educational objectives' taxonomy (Dell & Wantuch, 2017; Nash & Krauss, 2015), suitable for the needs of a particular academic subject (Darwazeh & Branch, 2015). To develop a high-quality instrument involved in the given investigation, the department incorporated the findings of the research studies and best practices described above.

The Development of the Web-Based Exams as Study's Instruments

The introductory statistics' instructors utilize an inquiry-based curriculum, which is focused on statistical thinking and interpretation through real data context (Makar & Ben-Zvi, 2011). In alignment with the inquiry-based approach, all assessments of the course have a theme, incorporate real data, and focus on statistical reasoning and interpretation. The statistics instructors identified the topics covered by each exam, corresponding course objectives, and the desired cognitive level of competencies students were expected to achieve by the end of each unit. Twenty-three automatically-scored questions were developed for each of four 70-minute exams. 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, according to which three minutes per higher order thinking item are sufficient to complete the assessment without rushing (Ladyshewsky, 2015). Multiple-choice and drop-down types of questions were selected to measure statistical reasoning and interpretation; free-response items, frequently more preferable type for measuring of higher order thinking, was not chosen due to an inability of the college LMS to assess these items automatically. For questions, answers to which require calculation, inserting one word or phrase, the instructors decided to use short-answer format. To reduce opportunities to guess and increase validity and reliability of exams, the number of multiple-choice and drop-down items was kept as small as possible. All exams have three drop-down, four multiple-choice, and 16 short answer questions.

To make sure that each exam item was aligned with needed cognitive processes and knowledge dimensions, the department first used the revised Bloom's taxonomy (Anderson et al., 2001) and subdivision into three major categories, knowledge-recall, interpretation-application, and problem solving-evaluation, mentioned by Dell and Wantuch (2017). However, the instructors realized that the inquiry-based approach with emphasis on statistical reasoning, critical thinking, and interpretation and learning outcomes of the course require more detailed taxonomy classification done by Darwazeh and Branch (2015). Some introductory statistics course objectives require analyzing followed by organizing and applying cognitive processes clearly described by Darwazeh and Branch. For example, to use an appropriate inferential technique, students have to analyze a question, then organize needed steps, and then apply the appropriate formula.

Like Nash & Krauss (2015), the department created a matrix consisting of Darwazeh & Branch's (2015) four knowledge and eight cognitive process dimensions. The difficulty of knowledge dimension increases from A to D; the difficulty of the cognitive processes increases from 1 to 8. The ninth dimension, meta-cognition, was not included because none of the objectives covered by exams required this cognitive process. As shown in Figure 2, each entry in the matrix has two components: the type of knowledge and the corresponding cognitive level. For example, the entry B6 indicates that cognitive application process occurs in the conceptual knowledge dimension.

The Knowledge Dimension	The Cognitive Process Dimension							
	1.Facts’ remembrance	2.Generalities remembrance	3. Comprehensio	4. Analyzing	5. Organizing	6. Application	7. Evaluation	8. Creation
A. Factual knowledge	A1	A2	A3	A4	A5	A6	A7	A8
B. Conceptual knowledge	B1	B2	B3	B4	B5	B6	B7	B8
C. Principles knowledge	C1	C2	C3	C4	C5	C6	C7	C8
D. Procedural knowledge	D1	D2	D3	D4	D5	D6	D7	D8

Figure 2. Knowledge and cognitive dimensions matrix.

The course objectives, knowledge and cognitive dimensions, question type, and number of points to be awarded were assigned to each exam question, as illustrated with three examples in Figure 3. None of the 23 exams’ items corresponded to fact remembrance, the lowest level of Darwazeh and Branch’ (2015) taxonomy. Three instructors, independently from each other, scrutinized the content of each question and the components of Figure 3 to make sure that the alignment between the objectives, cognitive processes, type of knowledge, and items’ content was done correctly.

Item	Learning Objective	Knowledge & Cognitive dimension	Question Type	Points
1	Identify sampling procedures	B4-D4, B6-D6, B7-D7	Drop-down	4
2	Calculate the sample proportion	A2-A5	Short-answer	1
3	Identify how to decrease the margin of error	B4-D4	Multiple-choice	2

Figure 3. The alignment of objectives, knowledge-cognitive dimensions, and items.

Several changes and improvements were made until the consensus between the three instructors was achieved.

In the next step, the instructors made sure that the exam questions have correct grammar, punctuation and spelling, include only relevant information, are positively stated, independent from each other, and not over-wordy. The experts also examined whether the response options to each multiple-choice or drop-down item are plausible, stated in the same verb tense, free from logical clues, have the same lengths, parallel structure, and only one correct answer. The developers of the exams included three to five response options in the drop-down and multiple-choice questions and did not include “none-of-the-above” or “all of the above” choices. The face and content validity of the exams were established. The reliability and construct validity of the exams are analyzed in Chapter 5.

To compare students’ performance on proctored and unproctored exams, the department created alternative forms of the same exam within each set. Thus, the tests within each set had the same questions, but with different themes based on different data sets. The use of alternative forms, in addition to randomization of the test items, reduces possible testing effects (Benedict & Zgaljardic, 1998; Feinberg et al., 2015) and increases validity and reliability of the exams (Towns, 2014). Moreover, the students did not know in advance that the exams in each set were parallel versions of the same test. All exams involved in the study were parts of a regular educational practice. For this reason, about two thirds of the questions in the second set covered slightly different areas in the curriculum than the questions in the first set. Inferential procedures with proportions were tested in the first set, while inferential procedures with means were assessed in the second set. The instructors designed the exams between the sets to be equivalent. The 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, and the same security mechanisms. Because each exam administered by the department included cumulative topics, six out of 23 questions were the same across all four exams. The same number of points was assigned to the corresponding questions in all four exams.

After the exams were developed and scrutinized for quality, they were piloted in the hybrid and online sections over the period of two semesters, and the consistency of scores and students' responses were analyzed. A similar pattern of scores had been observed during those semesters. Thus, the repeated use of the instruments yielded similar results, which established a tentative reliability of the tests (Frisbie, 1988). Although the instructors did not perform statistical item discrimination analysis, they observed that the exams discriminated low-achieving students from high-achieving students. The questions that were found by the students to be unclear or ambiguous were improved. Moreover, the instructors performed the item difficulty analysis: all the items that were too hard to solve and the majority of the students got them wrong, or were too easy because almost every student got it correctly, were revised. The scores on the improved exams correlated with the course grade, which indicated predictive validity (Frankfort-Nachmias & Nachmias, 2008). While the department has done some item analysis to improve the exams, reliability and construct validity of the instruments were not assessed statistically. For this reason, I performed reliability and construct validity analyses of the exams and reported the results in Chapter 4. In Fall 2015, the administration of the revised and improved web-based exams was extended to face-to-face sections.

The study's exams, developed in accordance with the best practices and research, satisfied all earlier mentioned Van Der Vleuten's (1996) criteria for high-quality assessment: feasibility, cost efficiency, acceptability by all stakeholders, educational impact, and adequate reliability and validity. The web-based exams utilized in the study were suitable for the introductory statistics course delivered through the LMC with classes scheduled in a computer classroom. All exams' questions were automatically-graded by the LMS, which reduced the cost associated with grading and saved faculty time. All instructors who teach web-assisted Introductory Statistics fully accepted the innovation by realizing the benefits of automatic grading, ability to administer a test online and use freed up in-class time for problem solving and projects, and availability of quick automatic item analysis provided by the LMS. The students appreciate the comfort of being able to take an exam at home. Thus, the acceptance of web-based exams by all stakeholders was achieved. Moreover, the exams include mostly higher order thinking questions. When students know that they have a test that requires higher cognitive processes, they are more actively involved in the learning process and preparations for exams (Jensen et al., 2014). Therefore, the exams had an educational impact. Lastly, content validity of the exams was established by the instructors, experts in the discipline; high reliability and constructed validity were confirmed during statistical analyses described in Chapter 4.

The two sets of web-based exams were adequate instruments to answer the study’s research questions. The scores obtained by the students on valid, reliable, and equivalent proctored and unproctored exams were used to test the format effect in each set. The administration of the instruments in the reverse order in the second set allowed for testing the order effect. The same instruments were used in all web-based sections taught by seven different instructors, which made possible to test the course delivery mode and instructor effects.

Manipulation of Independent Variables

Two sets of secured, web-based exams constituted the intervention of the study. The format and order in which the web-based exams were administered were manipulated. In the first set, the proctored exam was followed by the unproctored; in the second set, the order was reversed. The instructors of the department, experts in the subject matter, developed, piloted, and implemented the exams.

Operationalization of the Variables

The main independent variable of the study is the web-based exam format (IV1), proctored versus unproctored. Although proctored exam can be perceived as a cheating prevention mechanism, in the given investigation proctored is the value of the variable exam format. The proctored exams took placed in the computer classroom, while the unproctored exams could be completed at any location with the Internet accesses. However, the students were advised to take the unproctored exams in a quiet environment free of any distraction.

The exam score (DV) is the dependent variable. There were three drop-down, four multiple-choice, and 16 short answer items on each 50-point test. Each drop-down question was worth four points because there were four matching responses; students could earn partial credit for each correct response. Each multiple-choice question was worth two points; no partial credit was awarded for these items. Nine out of sixteen short-answer questions did not require calculation and were worth one point each without partial credit. For the remaining seven items, which were worth three points each, calculation was required, and students could earn partial credit on these questions. The answers to these questions had several parts. For example, the students were asked to represent the confidence interval as a three-part inequality with the appropriate symbol between the endpoints. One point was assigned for each correct endpoint and the symbol. The college LMS calculated the overall exam score in percent based on the points assigned to each question (Moodle, 2015). The samples of short-answer, drop-down, and multiple-choice exam items are provided in Appendix C, Appendix D, and Appendix E respectively.

The order (IV2) in which the proctored and unproctored exams were administered is the second independent variable. In the first set of two exams, the proctored exam was followed by unproctored; in the second set, the order of the exams was reversed. As explained previously, this choice of the order of the exams was dictated by the course curriculum, calendar, and faculty preferences in improving learning and instruction. The course delivery mode (IV3) is the third independent variable. This variable has three categories: web-assisted face-to-face, hybrid, and online. A web-assisted, face-to-face course is a course in which instruction occurs during regular class meetings, but up to 29% on the content is delivered through learning management system or other means of web-based technology (ITC, 2016). A hybrid course is a course with mandatory on-campus meetings that involve instruction; up to 79% of content is delivered through learning management system or other means of web-based technology. A fully-online course is a course that does not have mandatory face-to-face meetings that involve instruction; all content is delivered online (ITC, 2016). The instructor (IV4) is the last independent variable. There were seven statistics instructors whose students' scores were analyzed in the study.

Data Analysis Plan

The data analyses were conducted with the Statistical Package for the Social Sciences (SPSS) Version 23. When the data spreadsheet with all identifying information removed was obtained, the values of all variables involved in the study were carefully examined for missing, miscoded, or abnormal entries. After the screening, the data were imported into SPSS and examined for normality, possible outliers, heterogeneity of variance, and other assumptions of the study's statistical tests repeated and mixed ANOVAs.

The goal of the given investigation was to analyze the proctored and unproctored web-based exams' scores of community college students in web-assisted face-to-face, hybrid, and fully-online introductory statistics courses taught by a team of instructors who utilize the same curriculum and assessments. In the study, one group of students took two pairs (sets) of proctored and unproctored web-based exams. In the first set, the proctored exam was followed 7-10 days later by the unproctored exam; in the second set, the unproctored exam was followed 7-10 days later by the proctored one. I used the data archived by the department to answer the following research questions:

When equivalent automatically-scored web-based exams with the same security mechanisms are used,

RQ1: Is there a relationship between the exam format (IV1), proctored versus unproctored, and student scores (DV)?

> $H_0$1: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms.
>
> H_A2: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms.

RQ2: Is there a relationship between the order (IV2) in which proctored and unproctored exams are administered and student scores (DV)?

> $H_0$2***:*** There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the order in which exams are administered.

H_A2: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the order in which exams are administered.

RQ3: Is there a relationship between the course delivery mode (IV3), (a) web-assisted face-to-face, (b) hybrid, (c) fully online, and students' scores (DV)?

$H_0$3: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the course delivery mode.

H_A3: There is a significant difference in students' performance on automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the course delivery mode.

RQ4: Is there a relationship between the instructor (IV4) and students' scores (DV)?

$H_0$4: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the instructor of the course.

H_A4: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the instructor of the course.

At the college where the study took place, students can drop can drop classes until a certain day. Because the second pair of web-based exams was administered after the drop day, the number of students who took the first set of exams was different from the number of students who take both sets. For this reason, to test the exam format effect, a one-way repeated-measures ANOVA was applied to Set 1 exams' scores and Set 2 exams' score separately. A repeated-measures ANOVA is suitable for designs in which the dependent variable is repeatedly measured across all levels of one independent variable (Field, 2013; Mertler & Vannatta, 2002). In the given study, the exams' scores (DV) of each student were repeatedly measured across both levels of the independent variable exam format (IV1), proctored versus unproctored. Therefore, a repeated-measures ANOVA is an appropriate statistical test the format effect in Set 1 and Set 2.

In the further analysis, a mixed ANOVA was applied to test the order (IV2), course delivery mode (IV3), and instructor (IV4) effects. A mixed ANOVA is used when the dependent variable is measured for each level of the within-subject variable and each level of the between-subject variable (Field, 2013). Two within-subject variables, the format and order, and two between-subject variables, the course delivery mode and instructor, are involved in the given investigation. The variable format has two levels, proctored versus unproctored; the variable order also has two levels, proctored-unproctored order versus unproctored-proctored order. The variable course delivery mode has three levels: face-to-face, hybrid, and online. The scores of the students taught by seven instructors were analyzed; thus, the variable instructor has seven levels. The dependent variable exam score is measured for each level of the independent variables described above. Therefore, a mixed ANOVA is appropriate for testing of the order, course delivery mode, and instructor effects. A post hoc test was performed to determine where the difference occurred in significant results. A mixed ANOVA also allowed for testing interaction effects between all involved variables.

An independent ANOVA was used to compare the students across the course delivery modes and instructors with respect to GPA and age. The Principal Factor Analysis, Confirmatory Factor Analysis, and reliability analysis were performed to test the validity and reliability of the instruments. To use the statistical techniques, the corresponding assumptions should be verified (Field, 2013). I used box-plots to detect outliers and Shapiro-Wilk test to examine whether the dependent variable exam score was normally distributed across each category of each independent variable. Levene’s test was run to test homogeneity of variances. More detailed description of all statistical methods involved in the study is provided in Chapter 4.

Procedures to account for multiple statistical tests. To answer the study’s four research questions and test the corresponding hypotheses, several statistical tests were needed. Multiple testing conducted on the same data may increase the probability of making Type I error, which is commonly called familywise Type I error rate (Field, 2013). Field (2013) recommended using Bonferroni correction which adjusts the significance level for each statistical procedure such that the overall rate of Type I error across all tests stays less than .05. I used Bonferroni test for pair-wise comparisons.

Additional variables. In addition to the main independent variable exam format, the independent variables order, course delivery mode, and instructor were included in the analysis. As explained in detail in Chapter 2, these three variables were identified as additional variables because they could influence exam scores and bring threats to internal validity. To describe the study's population, student GPA, age, gender, major, academic units, number of working hours, and college grade level were utilized. To interpret the results of the study, I also compared the students in the course delivery mode and instructor groups with respect to GPA and age.

Reporting the results. After the appropriate statistical tests were applied, their results were reported in Chapter 4 and interpreted in Chapter 5. The measures of descriptive statistics, the means and standard deviations of exams' scores, were stated. For the repeated-measures ANOVA, I reported the F-ratio with the corresponding degrees of freedom, the p-value, the effect size, and the confidence interval as appropriate. For the mixed ANOVA, the F-ratios, p-values, and effect sizes for interaction effects between the involved variables was reported and interpreted as well. Cronbach's alpha coefficients were reported and interpreted for the reliability analysis; fit model indices were reported for Confirmatory Factor Analysis.

Threats to Validity

The main goal of the study was to determine whether there is a relation between the format of secured, web-based exams, proctored versus unproctored, and exams' score. A quasi-experimental one-group sequential design was chosen to examine this relation. However, quasi-experiments do not employ random assignment into groups, which can bring threats to external and internal validities (Shadish et al., 2002). The description of the threats to the study’s external and internal validities and ways of addressing them is provided below.

Threats to Study’s External Validity

Threats to external validity are factors that reduce generalizability of findings to other populations, settings, and times (Shadish et al., 2002). To increase external validity, Cook and Campbell (1979) recommended obtaining a heterogeneous representative sample and conduct an experiment in an environment that resembles a natural setting. Introductory Statistics students who took web-assisted face-to-face, hybrid, and online sections of the course constituted the target population. Because the scores of each individual in the target population were analyzed, the study’s sample was a census sample, which is representative by its nature. Moreover, the introductory statistics course is required for 80% of all transfer majors offered by the college. Thus, the study’s census sample was heterogeneous with respect to majors. Additionally, the study’s design elements are not related to the subject matter, which means that the used approach can be applied to any subject. Lastly, the web-based exams were administered in a natural educational setting as a part of regular educational practices. Therefore, the findings of the given study can be generalized to any higher education institution with a similar population of students.

Threats to Study’s Internal Validity

Threats to internal validity are factors that can influence a relationship between the main independent and dependent variables (Shadish et al., 2002). Selection, repeating testing, instrumentation, attrition, history, and maturation factors (Shadish et al., 2002) could constitute threats to the internal validity of the given study. Selection bias occurs when experimental and control groups are not equivalent, and the treatment effect is contaminated by the individual differences between the participants (Shadish et. al, 2002). During the first stage of the analysis, when the exam format effect was tested, each student was used as his or her own control, which rules out selection bias (Thompson & Panacek, 2006). However, at the college where the study took place, the students are not randomly enrolled in their classes, which could result in differences between the groups with respect to the course delivery mode and instructor. For this reason, the course delivery mode and instructor were included in the analysis as additional independent variables and their effects on the exam scores were tested. Additionally, the students' GPA and age were compared across the groups.

Testing bias. The testing bias is present when participants are tested more than once, and this repeated testing contaminates the treatment effect (Cook & Campbell, 1979; Shadish et al., 2002). Both types of testing bias, practice effect and fatigue effect, can be especially strong in a within-subject design where each student takes all tests included in the study (Shadish et al., 2002). In the given investigation, the use of alternative forms within each set, randomization of test items and response options, and long test-retest intervals of 7-10 days within the sets and 1 month between the set reduced threats to internal validity due to practice and fatigue effects. In addition to the testing bias, the order in which the tests are administered can influence exam scores (Shadish et al., 2002). For this reason, the effect of the variable order on the exams' scores also was tested.

Instrumentation bias. Instrumentation bias takes place when a measuring instrument changes over time (Shadish et al., 2002). The measuring instruments of the given investigation are four web-based exams. All four web-based exams, their content, wording, grading scale, question types, security mechanisms, and the LMS through which the exams were delivered, have been kept unchanged by the department from Fall 2015 to Summer 2017.

Attrition bias. Attrition, also known as experimental mortality, occurs when not all participants complete all stages of the study (Shadish et al., 2002). In the given study, the number of students in the first set was different from the number of students in the second set because several students who dropped the course before the exams in the second set were administered. Thus, the attrition bias could contaminate the manifestation of the independent variable when the exams scores of the second set were added to the analysis. There is no effective way to control or prevent attrition bias especially when archival data are used, but it can be tested (Miller & Hollist, 2007; Shadish et al., 2002). To test whether attrition bias was present in the study, I requested from the institution the dropout rate of the introductory statistics students during the 2014-2015 academic year, before the web-based exams were implemented, and during the 2015-2017 academic years, after the implementation took place. The results of the comparison of the attrition rates during these time frames are provided in Chapter 4.

History bias. History bias takes place when an event not related to the treatments occurs between different stages of an experiment and may contaminate the study's outcomes (Cook & Campbell, 1979; Shadish et al., 2002). The history threats can be reduced if the groups are selected from the same location, and the treatments are administered at the same time (Shadish et al., 2002). In the given study, all participants attended the same college, had the same curriculum, and used the same materials. Moreover, the students in all groups took all web-based exams at the same time. Therefore, a possible history threat to internal validity of the given study is minimized.

Maturation bias. Participants can become more mature during an experiment, which may threaten internal validity of the study if maturation influences the outcomes (Cook & Campbell, 1979; Shadish et al. 2002). To reduce maturation bias, Shadish et al. (2002) recommended involving participants of the same age from the same location. Most participants of the given investigation were young adults of the same age and lived close to the college. Therefore, it can be assumed that, on average, they matured and learned at the same pace.

Threats to Study's Construct Validity

Construct validity of a study is related to clear representations of involved constructs and assessing them (Shadish et al., 2002). Construct validity reflects whether the implemented intervention is the intervention that was intended to be implemented, and whether the outcomes were measured as it wanted to be measured (Trochim, Donnelly, & Arora, 2016). Construct validity allows for making inferences from a study's sampling particulars to the higher-order constructs these particulars represent

(Shadish et al., 2002). Inadequate explanation of constructs, mono-operational bias, and mono-method bias (Shadish et al., 2002; Trochim & Donnelly, 2006) are threats to constructed validity that can be related to the given investigation. To reduce the threats in regards with an inadequate explanation of constructs, each construct involved in the study was clearly defined operationally; quality of each operational definition was scrutinized.

Studies in which each construct is operationalized at multiple instances have higher construct validity than mono-operational investigations (Shadish et al., 2002). If a specific intervention is implemented in one class at one point in time, the full breadth of the concept of the intervention may not be captured (Trochim & Donnelly, 2006). In the given study, the intervention, the secured, web-based exams, are administered in many sections delivered through different course modes taught by seven different instructors during several semesters. This fact decreases mono-operation bias and increases construct validity of the investigation.

Mono-method bias occurs when there is only a single version of a measure presented in one way (Shadish et al., 2002; Trochim & Donnelly, 2006). To reduce mono-method bias, Trochim & Donnelly (2006) recommended implementing of multiple measures of main constructs and verifying through piloting that the used measures assess correctly. There are four exams in the given study, which included drop-down, multiple-choice, and short-answer items. These exams were piloted, evaluated, and improved during two consecutive semesters. Therefore, the mono-method bias is minimized.

Threats to Study’s Statistical Conclusion Validity

Statistical conclusion validity reflects whether the existence of the relationship between the independent and dependent variables and the strengths of this relation were inferred correctly (Shadish et al., 2002; Trochim & Donnelly, 2006). Shadish et al. (2002) identified nine threats to statistical conclusion validity. The description of these threats and how they were addressed in the given investigation is provided below.

Low statistical power. The power of a test is the probability that the test will reject the null hypothesis when the hypothesis is false (Shadish et al., 2002). Thus, power measures the test ability to detect the relationship between the independent and dependent variables correctly. Low power may result in an incorrect conclusion that there is no relationship between the independent and dependent variables. To increase power, Shadish et al. (2002) recommended using a larger sample size, utilizing several stages of measurement, and incorporating a within-subject design with the control for testing and order effects. A big population size of 850 students was analyzed in the given investigation. Moreover, there were two stages of measurements: the students were tested-retested in the middle and at the end of each semester. Lastly, a within-subject design was utilized to test the format and order effects. All these aspects increased the study’s power.

Violated Assumptions of Statistical Tests. Violations of assumptions of statistical tests involved in a study bring potential threats to conclusion validity (Shadish et al., 2002). Normality and homogeneity of variances were assumptions of the study’s statistical tests, a repeated ANOVA and mixed ANOVA. If the sample size is large, as it is in the given study, ANOVA is robust to nonnormality (Field, 2013; Shadish et al., 2002). The analyses of all statistical assumptions in described in detail in Chapter 4.

Error Rate Problem. As previously mentioned, multiple statistical testing conducted on the same data set can increase familywise Type I error rate (Field, 2013). ANOVA in SPSS automatically corrects for some error rate problems occurring in multiple comparison tests (Field, 2013). Reporting a combination of effect sizes, confidence intervals, and p-values can reduce error rate problem and increase conclusion validity (Shadish et al., 2002). I reported the confidence intervals, p-values, and effect sizes in all study's tests where it was appropriate.

Unreliability of Measures. An inference about a relationship between the independent and dependent variables may be inaccurate if each of the variables is measured unreliably (Shadish et al., 2002; Trochim & Donnelly, 2006). For this reason, reliability of each measurement should be evaluated and reported (Shadish et al., 2002). In the given investigation, a lot of effort has been put into measuring the main independent variable exam format, proctored versus unproctored, and the dependent variable exam scores as accurately as possible. The web-based exams administered in class had been carefully proctored by faculty. The unproctored exams could be taken at any location with the Internet access, but the students were asked to complete the tests in a quiet environment. As previously described in the Instrumentation section, the secured equivalent exams were developed in accordance with the best practices; their tentative reliability was established by the instructors, experts in the subject matter. To evaluate the reliability of all exams statistically, I performed Cronbach’s alpha reliability analysis. This analysis is described in detail in Chapter 4.

Restriction of Range. The power of a study and the relationship between involved factors can be influenced by restricted variables. There were no cut scores or other restrictions on the dependent variable in the given investigation. All students’ score, regardless of their values, were analyzed.

Unreliability of Treatment Implementation. Inconsistency in treatment implementation can be a threat to the conclusion validity of a study (Shadish et al., 2002; Trochim & Donnelly, 2006). In the given investigation, all involved instructors were asked to follow the same procedure in administering the web-based exams. All students went through the same steps in accessing the exams and submitting their responses. The exams were delivered through the same LMS, which automatically graded and recorded students’ scores. All these steps reduced the unreliability of treatment implementation. However, during the data screening, it was found that about 50 students took both exams in Set 2 in unproctored format. The scores of these 50 students were analyzed separately in the Additional Statistical Tests section in Chapter 4.

Extraneous Variance in the Experimental Setting. The treatment effects can be measured inaccurately if there are extraneous factors in an experimental setting (Shadish et al., 2002; Trochim & Donnelly, 2006). Time of day or a distracting noise may be such extraneous variance (Trochim & Donnelly, 2006). Synchronous administering of the unproctored exams in the given study eliminated extraneous factor time of day. Moreover, all students were asked to take the unproctored exams in a quiet environment.

However, there is no way to control whether the participants are not distracted during unsupervised tests. On the other hand, distraction during unproctored tests may be present in any unsupervised environment. Therefore, in spite of inability to control for destruction during unsupervised testing, the results of the study can be generalized for similar secured unproctored web-based exams administered in similar institutions.

Heterogeneity of Units. The threat related to heterogeneity of units occurs when the participants are very diverse and widely vary on the measures of the dependent variable (Shadish et al., 2002; Trochim & Donnelly, 2006). I utilized a within-subject approach in testing the format and order effects, which eliminate this threat. The between-subject factors, the course delivery mode and instructor, could bring some differences, but these factors were phenomena under investigation, and their impact on exams' scores was studied.

Inaccurate Effect Size Estimation. An effect size is a standardized measure of the magnitude of an effect or the strengths of a relationship between the variables (Field, 2013; Murphy et al., 2014). Inaccurately estimated effect size can be another threat to the statistical conclusion validity (Murphy et al., 2014; Shadish et al., 2002). Theory of research and previous studies on the given topic can provide estimation for the effect size (Murphy et al., 2014). If there is not enough information to estimate the effect, power analysis can be performed by utilizing a conservative estimate. An investigation with adequate power to reliably identify small or medium effects also can have enough power to detect large effects. If a study is designed with a large effect in mind, there might be insufficient power to detect small but crucial effects (Murphy et al., 2014). SPSS calculates effect sizes for all ANOVAs used in the study. I reported the effect sizes in Chapter 4.

Ethical Procedures

The following archival data were requested from the department: the scores obtained by the students on all web-based exams involved in the study, exam format, instructor, semester, course delivery mode and some demographics such as GPA, age, gender, number of units, and major. The Institutional Review Board (IRB) at the college where the study took place permitted me to conduct the study, get access to the data, and use them for analyses and reports. The Walden University IRB approval (05-23-17-0315459) was obtained on May 23, 2017.

Because the archival data collected in a natural educational setting were used in the given investigation, there were no actual recruitment or selection of the participants, and no consent forms were necessary. However, the data provided by the department satisfied all ethical procedures' requirements of the IRB of the college where the study took place and the IRB at Walden University. To protect the involved individuals, all identification information of the students and instructors was removed from the spreadsheet and recoded by the Introductory Statistics coordinator. Thus, I could not associate the information included in the data set with any of the students or their instructors.

After the Walden IRB approved my application, I requested the data from the department. The Introductory Statistics coordinator provided the anonymous and coded electronic spreadsheet with all requested information needed for the study. I copied the spreadsheet to my computer and on a memory stick for backup. My computer is password-protected; the memory stick is kept locked in my office. I am the only one who has access to the data, which will be kept for at least five years and then deleted from the computer and memory stick.

I am one of the instructors who teach Introductory Statistics in our department. However, the removal of identifying information did not allow me as a researcher to know which data correspond to my students. Moreover, I do not supervise the departmental programs, faculty, and data collection process. Thus, there was no conflict of interest or power differential.

Summary

In the given quantitative study, a quasi-experimental one-group sequential design was utilized to compare individual student scores on two sets of equivalent automatically-scored web-based exams in community college face-to-face, hybrid, and online Introductory Statistics sections. The sections were taught by seven instructors who used the same curriculum and assessments. The secured exams were administered in a certain sequence. In the first set, the proctored exam was followed 7-10 days later by the unproctored exam, and, in the second set, unproctored exam was followed 7-10 days later by the proctored one. The archived scores obtained by the students on these exams were requested from the department and analyzed; all students were perceived as one group. To answer the first research question whether there is a relationship between the format in which exams are administered (IV1), proctored versus unproctored, and test score (DV), a one-way repeated-measures ANOVA was applied separately to Set 1 and Set 2 exams’ scores. A mixed ANOVA was utilized to answer other three research questions by testing the order (IV2), course delivery mode (IV3) and instructor effects (IV4).

A detailed description and analysis of the obtained data and demographics of the participants is provided in Chapter 4. The statistical tests’ results are analyzed and reported in this chapter as well. A complete discussion of the answers to the study’s research questions concludes Chapter 4.

Chapter 4: Results

The purpose of this quantitative study was to investigate whether inconvenient and expensive proctoring is necessary when web-based exams with systematically selected nonbiometric security mechanisms are used. The relationship between the format in which equivalent automatically-scored secured, web-based exams were administered, proctored versus unproctored, and exam scores was examined. Additionally, the order, course delivery mode, and instructors’ effects were analyzed to answer the following research questions.

When equivalent automatically-scored web-based exams with the same security mechanisms are used,

RQ1: Is there a relationship between the exam format (IV1), proctored versus unproctored, and student scores (DV)?

> $H_0$1: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms.
>
> H_A1: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms.

RQ2: Is there a relationship between the order (IV2) in which proctored and unproctored exams are administered and student scores (DV)?

$H_0$2: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the order in which exams are administered.

H_A2: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the order in which exams are administered.

RQ3: Is there a relationship between the course delivery mode (IV3), (a) web-assisted face-to-face, (b) hybrid, (c) fully online, and students' scores (DV)?

$H_0$3: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the course delivery mode.

H_A3: There is a significant difference in students' performance on automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the course delivery mode.

RQ4: Is there a relationship between the instructor (IV4) and students' scores (DV)?

$H_0$4: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the instructor of the course.

H_A4: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the instructor of the course.

At the beginning of this chapter, I describe the period for data collection and participation rate, provide demographic characteristics of the population, and discuss challenges associated with the implementation of the intervention. Detailed data analysis and findings are provided in the main part of the chapter. I conclude the chapter with the summary of the answers to the study’s research questions.

Data Collection

To answer the study’s research questions and test the hypotheses, I utilized the archival scores obtained by the introductory statistics students on two sets of secured proctored and unproctored web-based exams, Set 1 and Set 2. These exams were implemented and administered by the department as a part of regular educational practices. In the first set, which took place in the middle of each semester, the proctored web-based exam was followed by the unproctored one. In the second set, which was administered at the end of each term, the order was reversed.

Time Frame for Data Collection

After I received Walden University IRB approval, I requested and obtained the exam scores and demographics of all students enrolled in the web-based statistics sections offered in face-to-face, hybrid, and online formats during the Fall 2015 through Summer 2017 semesters. Additionally, the institution where the study was conducted provided the data on attrition rate of the statistics students during the Fall 2014-Spring 2015 semesters, before the web-based exams were implemented, and the Fall 2015-Summer 2017 semesters, after the implementation took place. I used these attrition rate data to test whether attrition bias due to dropout took place. Because the archival data collected in a natural educational setting were used, no actual recruitment took place.

Discrepancies in Data Collection

There were no discrepancies in data collection from the plan described in Chapter 3. When the IRB approval was obtained, I requested the archival exam scores, demographics, and attrition rate data archived by the college. The needed information was provided by the institution in coded spreadsheets in electronic form.

Overall Participation Rate

According to the data provided by the institution, 1,150 students were enrolled in 33 web-assisted introductory statistics sections during the study's time frame. Out of these 1,150, 850 students (74%) took at least one exam involved in the study. Although there were students who did not take the study's first exam but took the second one, or did not take the first two exams, but took the last two exams, the number of examinees was decreasing with each test due to drop out. The total number of dropped students was 366. Thus, the attrition rate constituted about 31.8 %. The analysis of whether the attrition rate during the study was significantly different from the attrition rate before the web-exams were implemented is provided in the Additional Statistical Tests section.

Participation Rates on Set 1 and Set 2 Exams

In the first pair of the exams involved in the study, the proctored exam followed the unproctored one. Out of 850 students who took at least one study's exam, 838 completed Set 1 proctored exam and 807 completed Set 1 unproctored exam. On Set 1 proctored exam, 28 students had extended test time in accordance with their approved accommodations. All these 28 students were enrolled in face-to-face courses. On Set 1 unproctored exam, the number of students with extended test time was 37. Fifty-four of the 807 students could not take the unproctored exam synchronously with all other students and took the parallel version of the exam at different time, also in unproctored format. Two of the 54 students had extended test time. The number of students who took both Set 1 exams was 776. Out of these 766, 52 took the parallel version of the unproctored exam, and 16 students had extended test time.

In Set 2, the unproctored exam, which was administered first, was completed by 737 students, 86 of whom had a schedule conflict with the synchronous administration of the test and took its parallel version at different time in unproctored format. Eighteen of 737 students had extended test time. The second exam in Set 2 was completed by 739 students. Although this exam had to be proctored, out of these 739 participants, 683 students took the exam in proctored format and 56 in unproctored. Thus, the 56 students took the first and the second exams in Set 2 in unproctored format. On Set 2 unproctored exam, 13 students had extended test time. The number of students who took both Set 2 exams was 718, 76 of whom took the parallel version of the unproctored test, 51 students took both Set 2 exams in unproctored format, and 13 students had the extended time accommodation. Seven hundred and two students completed all four exams. Out of these 702 students, 67 took the parallel version of at least one unproctored exam, 51 were the students who took both Set 2 exams in unproctored format, and 15 students completed at least one of the four exams with extended time. To examine the entire population under investigation, the scores of all 850 students who took at least one study's exam were used in the analyses, including the reliability and construct validity analyses of involved exams provided in the Additional Statistical Tests section.

Baseline Descriptive

The college where the study took place offers Introductory Statistics, a traditional 4-unit course that satisfies all requirements for university transfer, in web-assisted face-to-face, hybrid, and fully-online course delivery modes. Students in face-to-face classes, which are scheduled during the entire day, meet two times a week for 2 hours either on Monday and Wednesday or Tuesday and Thursday. Students in hybrid sections of the course come on campus on some Saturdays, while fully-online students do not have mandatory campus meetings that involve instructions.

Out of 33 sections involved in the study, 26 were offered in a face-to-face mode ($N_{students}$ =703), four sections were delivered as a hybrid ($N_{students}$ = 92), and two sections were fully-online ($N_{students}$ =55). Morning sections of the course ($N_{students}$ =132) were offered from 8 a. m. to 10 a.m., day sections ($N_{students}$ =233) from 11 a.m. to 1 p.m., afternoon sections ($N_{students}$ =221) from 2 p.m. to 4 p.m., and late afternoon sections ($N_{students}$=100) from 4:30 p.m. to 6:30 p.m. Additionally, there was one small evening section offered from 7 p. m. to 9 p. m. ($N_{students}$=17).

Students' Demographics

The participants' age ranged from 14 years to 50 years, with the mean of 22. The mean age of the face-to-face and online students was 22 years (M_{f2f} = 21.59, M_{onl} = 21.98), while the hybrid students constituted the older group with the average age of 25 years (M_{hyb}= 25.17). The mean GPA was 3.19 (M_{f2f} = 3.19, M_{hyb}= 3.13, M_{onl} = 3.23). Overall, there were 488 females (57%) and 362 males (43%). The study's population included 410 continuing students (48.2%), the students who attended the college for more than 1 year, 200 transfer students (23.5%), the students who after completion the introductory statistics course transferred to universities, and 142 freshmen (16.7%), the students who were the first-year community college students. There were also 37 graduate students (4.4%), the students who already had an undergraduate degree and took the course as a prerequisite for a graduate school admission, 36 high school students (4.3%), who completed the class concurrently with their high school courses, and 25 undergraduate students (2.9%), the students who took the course concurrently with their undergraduate programs' classes. About 70% of the graduate students were enrolled in the hybrid or online sections, while 60% of the undergraduate students and 81% of high school students completed the course in face-to-face mode.

The participants were majoring in business (27%), nursing (13%), psychology (10%), biology (5%), criminology (4.5%), economics (4%), communications (3.7%), kinesiology (3.7%), sociology (3%), math (1.5%), engineering (1.5%), film (.9%), statistics (.5%), anthropology (.5%), and computer science (.5%). Art, dietetics, and physics majors constituted less than .5%. The major of the remaining students (about

20%) was undecided. Over 67% of the participants had at least one job and, on average, worked for 24.2 hours a week. Out of these 67%, about 14 % worked up to 10 hours a week, 32% between 11 and 22 hours a week, 29%, between 21 and 30 hours a week, 21% between 31 and 40 hours a week, and 4% worked for more than 40 hours a week. The participants were taking from 4 to 26 academic units during the semester when they took the statistics course, with the mean of 12.7 units.

Instructors' Demographics

Seven Introductory Statistics instructors were involved in the study. Their experiences of teaching web-based introductory statistics courses ranged from one semester to 18 semesters, with the average of 9.3 semesters. The instructors' overall teaching experience at the college where the study took place ranged from one to 19 years, with the mean of 9.7 years. At the time when the web-based exams were implemented, the instructor's experience with web-based testing ranged from zero to five semesters. The age of the instructors ranged from 32 to 64 years with the mean of 42.6 years. All instructors were females. During the study's time frame, each instructor was teaching from one to three sections of the course per semester.

Representativeness of the Sample

Because the scores of the entire population of web-based introductory statistics students were analyzed, the census sample was utilized in the study. The census sample represents the entire population under investigation. Therefore, the study's findings can be generalized for institutions with similar populations.

Treatment and Intervention Fidelity

The department, the archival data of which I analyzed in the study, implemented secured proctored and unproctored web-based exams in web-assisted introductory statistics courses. This departmental implementation of web-based exams with systematically selected security mechanisms can be perceived as an intervention. The proctored format of the exams was the control condition, while the unproctored format was the experimental one. Each student went through both conditions by taking two proctored and two unproctored exams in the following sequence: Set 1 proctored, Set 1 unproctored, Set 2 unproctored, Set 2 proctored.

Overall, the administration of the web-based exams occurred as it was planned by the department and described in Chapter 3. Set 1 proctored web-based exam was administered in class in the middle of the semester and supervised by the instructors. Seven days after that, on Saturday, at 9:30 a.m., the students in all sections took Set 1 synchronous unproctored web-based exam off campus. Set 2 synchronous unproctored exam was administered to all students on another Saturday at the end of the semester, also off campus, at 9:30 a.m. A minor challenge was related to the fact that some students could not take the unproctored exams synchronously. Synchronous administration of the unproctored exams was an important security mechanism used by the department. To overcome this challenge, the department created one more parallel version of the web-based exam to accommodate the students with the schedule conflicts, and, at the same time, reduce possible dissemination of the exams' questions. Thus, if a student could not take an unproctored exam during the scheduled time because of serious reasons, he or she completed a parallel version of the web-based exam in an unproctored format at a different time as it was agreed with the instructor.

No technical difficulties or Internet problems were reported by the faculty and their students during proctored exams. A minor discrepancy was detected during the data screening process: In two sections of the course (*N* =56), both Set 2 exams were administered in unproctored format. There were no adverse effects with serious consequences related to the implementation of the secured, web-based exams.

Results

The format in which the web-based exams were administered, proctored versus unproctored, was the main independent variable of the study. The exam score was the dependent variable. The department implemented two pairs of web-based exams: in Set 1, the proctored exam was followed by the unproctored one, in Set 2, the order was reversed. Thus, there were four sets of scores: Set 1 proctored exam (S1PE) score, Set 1 unproctored exam (S1UPE) score, Set 2 unproctored exam (S2UPE) score, and Set 2 proctored exam (S2PE) score. The order in which the exams were administered, course delivery mode, and instructor were additional independent variables.

As I described previously, not all students in the entire population under investigation (total group) completed the exams with all security mechanisms. There were students who took both exams in each set with all security mechanisms utilized by the department (the main group), students who could not take the unproctored exams synchronously (V2 group), students who had extended test time (Ext. Time group), and students who took both exams in Set 2 in an unproctored format (UP group). The analyses of V2, Ext. Time, and UP groups in relation to the study's research questions are provided in Additional Statistical Analysis section. I organized the he main analysis by the research questions and corresponding hypotheses.

For each research question, I first analyzed the descriptive statistics of the involved variables for the total and main groups. Then, I narrowed the analyses to the main group only, verified the statistical assumptions, conducted the appropriate statistical tests, and reported the results. In the reports of these findings, I included exact *F* statistics, associated *p*-values, the 95% CI for the differences of scores, when it was appropriate, and the effect size η^2. According to Cohen (1992), $\eta^2 =.02$ corresponds to a small effect size, $\eta^2 =.13$ to a medium effect size, and $\eta^2 =.26$ to a large effect size. All effects were reported as significant at $\alpha =.05$. I began with the analyses of the first research question.

Testing the Hypotheses of RQ1

The first research question and the corresponding hypotheses were formulated in the following form:

When equivalent automatically-scored web-based exams with the same security mechanisms are used,

RQ1: Is there a relationship between the exam format (IV1), proctored versus unproctored, and student scores (DV)?

> $H_0$1: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored

proctored introductory statistics web-based exams with the same security mechanisms.

H_A2: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms.

In Set 1, 838 students took the proctored exam, S1PE, which was administered first. In this total group, the scores ranged from 0% to 100% with the mean of 66.43%. After I removed the scores of 28 students with the extended test time, the mean score of the main group slightly increased to 66.80%. The second exam in Set 1, S1UPE, the unproctored exam, was completed by 807 students. The mean score of these students was 67.80%. After I excluded the scores of the students who took version 2 of this exam (*N* =54), the mean score increased to 68.23%. The additional removal of 34 students with extended test time increased the mean further: the average score of the resulting main group was 68.72%. The descriptive statistics of the scores on S1PE and S1UPE for the total and main groups is provided in Table 1.

Table 1

Descriptive Statistics of S1PE and S1UPE Scores in Total and Main Groups

Measure	TG PE	MG PE	TG UPE	MG UPE	$Diff_{TG}$	$Diff_{MG}$
N	838	810	807	719		
M	66.43	66.84	67.80	68.72	-1.37	-1.88
SD	21.52	21.45	21.29	20.98	.23	.47

Note: TG PE = total group on S1PE; MG PE = main group on S1PE; TG UPE = total group on S1UPE; MG UPE = main group on S1UPE; $Diff_{TG}$ = difference in scores on S1PE and S1UP in the total group; $Diff_{MG}$ = difference in scores on S1PE and S1UP in the main group.

As seen in Table 1, the removal of the scores of the students who took version 2 of S1UPE and the students with extended test time changed the descriptive statistics on both exams by less than 1%. In both groups the students' scores on unproctored exam were higher than on proctored exams, but no more than by 2%; the SDs decreased by less than .5%. Next, I paired S1PE and S1UPE scores.

Both S1PE and S1UPE were taken by 795 students. After all exclusions, the number of students in the main group was 732. The descriptive statistics for the paired scores in Set 1 in the total and main groups is presented in Table 2.

Table 2

Descriptive Statistics of Paired Set 1 Scores in Total and Main Groups

Measure	TG PE	MG PE	TG UPE	MG UPE	*Diff*TG	*Diff*MG
N	795	732	795	732		
M	66.95	67.26	67.97	68.56	-1.02	-1.30
SD	21.09	21.18	21.16	21.12	-.09	.06

Note: TG PE = total group on S1PE; MG PE = main group on S1PE; TG UPE = total group on S1UPE; MG UPE = main group on S1UPE; *Diff*TG = difference in scores on S1PE and S1UP in the total group; *Diff*MG = difference in scores on S1PE and S1UP in the main group.

According to Table 2, after the pairing took place, the students' scores on the unproctored exam were higher than on the proctored one by about 1% in both groups. The SDs in the total and main groups changed by less than .1%. Further analysis was narrowed down to the main group only because all students in this group took both exams with all security mechanisms utilized by the department. To test whether the difference in scores on proctored and unproctored exams observed in the descriptive statistics analysis was significan, I applied a repeated measures ANOVA to Set 1 main group, but before I ran the analysis in SPSS, I examined whether the assumptions of a repeated measures ANOVA were met.

The validity of the results of a repeated measures ANOVA depends on three assumptions: (a) there is no dependency in the scores between participants, (b) the dependent variable score is normally distributed for each level of within-subject factor, (c) the population variance of difference scores computed between any two levels of a within-subject factor is the same (sphericity) (Field, 2013). The sphericity assumption is not applicable if there are only two levels of the independent variable, as it is in RQ1. ANOVA is robust to violations of the normality assumption if the sample size is at least 30 (Field, 2013).

In the given study, there was no dependency in scores between the participants on all exams. To determine whether the dependent variable score was normally distributed on proctored and unproctored exams in Set 1, I ran graphical and numerical analysis of normality in SPSS for the paired scores in Set 1 (N = 732). According to the boxplot, the distribution of the scores on S1PE was slightly left-skewed without outliers. Shapiro-Wilk test, a normality test in SPSS, was significant (p <.001), indicating deviations from normality. However, ANOVA is robust to nonnormality if the sample size is at least 30. The skewness of -.37 (*std. error* = .092) and kurtosis of -.81 (*std. error* = .183), were between -2 and 2, which is the acceptable range. The distribution of S1UPE scores was also left-skewed with three mild outliers of 4%, 6%, and 8%. Shapiro-Wiki test was significant ($p \leq$.004). The skewness of -.60 (*std. error* = .092) and kurtosis of -.31 (*std.*

error = .183), were in acceptable limits and could not affect the validity of the used statistical test. All assumptions of the ANOVA were met; I ran the test in SPSS.

Results of Testing the Null Hypotheses of RQ1 in Set 1

In Set 1, the proctored exam followed by the unproctored one. In the main group of students (*N* =732) who took both Set 1 exams with all security mechanisms utilized by the department, there was no significant difference between the scores on the equivalent web-based proctored and unproctored exams with the same security mechanisms ($F(1,731) = 2.38$, $p = .12$, $\eta^2 = .00$). The corresponding 95% confidence interval for the difference of S1PE and S1UPE scores included 0% (95% CI [-2.94, .35]), which supported the insignificant format effect. The null hypothesis of RQ1 in Set 1 was retained. Next, I repeated the analyses for Set 2 exam scores.

The unproctored exam S2UPE was administered first; 739 students completed this exam. Their scores ranged from 4% to 100% with the mean of 62.10%. After I excluded the scores of the students who took the second version of S2UPE and students who had extended test time, the mean score of the remaing 637 students became 63.12%.

The same number of students ,739, took the second exam in Set 2, which had to be proctored. The mean score of the 739 students was 68.49%. However, out of these 739 students, 56 students took the second exam in Set 2 in unproctored format. I removed the scores of these 56 students and the scores of 12 students with extended test time. The mean score of the remaining 671 students became 68.52%, which was very close to 68.49%, the mean of the initial group of 739 students. The descriptive statistics of the scores on S2UPE and S2PE in the total and main groups is shown in Table 3.

Table 3

Descriptive Statistics of S2PE and S2UPE Scores in Total and Main Groups

Measure	TG UPE	MG UPE	TG PE	MG PE	*Diff*TG	*Diff*MG
N	739	637	739	671		
M	62.10	63.12	68.49	68.52	-6.39	-5.40
SD	20.12	20.01	18.32	18.12	1.80	1.89

Note: TG UPE = total group on S2UPE; MG UPE = main group on S2UPE; TG PE = total group on S2PE; MG PE = main group on S2PE; *Diff*TG = difference in scores on S2PE and S2UP in the total group; *Diff*MG = difference in scores on S2PE and S2UP in the main group.

According to Table 3, the students in both groups had higher scores on the second, proctored exam, than on the first unproctored exam. In the total group, the scores increased by 6.4%; in the main group, the scores increased by 5.4%. While the mean scores increased from the first to the second test administration, the SDs in both groups decreased by slightly less than 2%. Next, I paired the scores of all students who took S2UPE and S2PE2.

In Set 2, 725 students completed both exams. After I removed the scores of the students who took the second version of S2UPE, students who had extended test time, and students who took both Set 2 exams in unproctored format, the scores of 593 students remained. The descriptive statistics of the paired scores in Set 2 is provided in Table 4.

Table 4

Descriptive Statistics of Paired Set 2 Scores in Total and Main Groups

Measure	TG UPE	MG UPE	TG PE	MG PE	*Diff*$_{TG}$	*Diff*$_{MG}$
N	725	593	725	593		
M	62.59	63.82	68.95	69.51	-6.36	-5.69
SD	19.87	19.87	18.00	17.79	-1.87	2.08

Note: TG UPE = total group on S2UPE; MG UPE = main group on S2UPE; TG PE = total group on S2PE; MG PE = main group on S2PE; *Diff*$_{TG}$ = difference in scores on S2PE and S2UP in the total group; *Diff*$_{MG}$ = difference in scores on S2PE and S2UP in the main group.

According to Table 4, the students in both groups had about 6% higher scores on Set 2 proctored exam than on the unproctored one. In the total group, the SD was about 2% higher on the unproctored exam; in the main group, the SD was about 2% higher on the proctored exam. Further analysis was restricted to the main group. To test whether the difference in scores on proctored and unproctored exams in Set 2 was significant, I verified statistical assumptions for a repeated measures ANOVA and applied the test.

According to the boxplots, the distributions of S2PE and S2UPE scores in the main group (N = 593) were slightly left-skewed. There was one outlier of 10% in S2PE and one outlier of 4 % in S2UPE. Shapiro-Wiki test was significant for the scores of both exams (p <.001). The skewness and kurtosis were in acceptable range (S2PE: *skewness* = -.34, *kurtosis* = -.47; S2UPE: *skewness* = -.31, *kurtosis* = -.52). The assumptions of the repeated measures ANOVA were met. I ran the test in SPSS.

Results of Testing the Null Hypothesis of RQ1 in Set 2

In Set 2, a medium significant format effect was observed (N =593, F(1, 592) = 101.44, $p \leq .004$, η^2 =.15). The corresponding confidence interval for the difference of scores excluded zero (95% CI [4.58, 6.80]), which supported the significant effect. The null hypothesis of RQ1 in Set 2 was rejected. For comparison, I combined the results of RQ1 analyses in both sets in Table 5.

Table 5

Results of RQ1 Analysis with Set 1 and Set 2 Scores

							95% CI	
Group	*N*	M_{PE}	M_{UE}	*F*	*p*	η^2	*LL*	*UL*
S1 Main Group	732	67.26	68.56	2.38	.12	.00	-2.94	.35
S2 Main Group	593	69.51	63.82	101.44	.00	.15	4.58	6.80

Note: M_{PE} = the mean score on proctored exam in %, M_{UP} = the mean score on unproctored exam in %; 95% CI = confidence interval for the differences between proctored and unproctored exam scores; *LL*= lower limit, *UP* = upper limit.

According to Table 5, in Set 1, the format effect was not significant. However, in Set 2, a medium significant format effect was observed. Interpretations of these results are given in Chapter 5. Next, I proceeded to RQ2 analyses.

Testing the Null Hypothesis of RQ2

The second research question and the corresponding hypotheses were formulated as follows:

When equivalent automatically-scored web-based exams with the same security mechanisms are used,

RQ2: Is there a relationship between the order (IV2) in which proctored and unproctored exams are administered and student scores (DV)?

> $H_0$2: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the order in which exams are administered.

> H_A2: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the order in which exams are administered.

Two independent variables were involved in RQ2: exam format and exam order. The exam score was the dependent variable. The student scores on all four exams were needed to test the null hypothesis of the second research question. For this reason, first I matched four scores for each student who took all study's exams and then proceeded to the descriptive statistics analysis of the involved scores.

There were 713 students who took all four web-based exams. In this total group the means were almost the same on S1PE, S1UPE, and S2PE, while the mean score on S2UPE was about 7% lower (M_{S1PE} = 70.48%, M_{S1UPE} = 70.89%, M_{S2PE} = 69.17%, M_{S2UPE} = 62.76%). After I removed the scores of 54 students who took both exams in Set 2 in unproctored format, the means stayed almost the same (M_{S1PE} = 70.57%, M_{S1UPE} = 70.62%, M_{S2PE} = 69.14%, M_{S2UPE} = 63.09%). Next, I excluded the scores of 107 students who took the second version of at least one unproctored exam, which slightly increased the mean scores on all four exams (M_{S1PE}= 71.27%, M_{S1UPE} = 71.57%, M_{S2PE} = 70.07%, M_{S2UPE} = 64.10%). The additional exclusion of 27 students who had extended time on at least one of the four exams did not change the means a lot (M_{S1PE}= 71.78%, M_{S1UPE} = 71.86%, M_{S2PE} = 70.22%, M_{S2UPE} = 64.24%). The summary of the descriptive statistics of the scores on all four exams in the total and main groups is shown in Table 6.

Table 6

Descriptive Statistics of all four Exam Scores in Total and Main Groups

M_{S2PE}(SD)	*N*	*M_{S1PE}(SD)*	*M_{S1UE}(SD)*	*M_{S2UE}(SD)*	*M_{S2PE}(SD)*
Total Group	713	70.48 (19.37)	70.89 (19.31)	62.76 (19.82)	69.17 (17.87)
Main Group	525	71.73 (19.14)	71.86 (19.08)	64.25 (19.92)	70.22 (17.56)

Note: *M_{S1PE}* = the mean score on Set 1 proctored exam in %; *M_{S1UE}* = the mean score on Set 1 unproctored exam in %; *M_{S2UE}* = the mean score on Set 2 unproctored exam in %; *M_{S2PE}* = the mean score on Set 2 proctored exam in %.

According to Table 6, when the proctored exam was administered first and unproctored exam was administered second the mean score on proctored exam was higher by 0.4 % in the total group and by 0.1% in the main group. When the unproctored exam was administered first and the proctored one was administered second, the mean score on the proctored exam was also higher, but with the difference in scores of 6.4% in the total group and 6.0% in the main group. Moreover, in both groups, when the unproctored exam was administered first, the mean score on this exam was lower by about 7% than the mean scores on other three exams. The difference in scores on the proctored exams was no more than 1.5%. This descriptive statistics analysis suggested that the order in which proctored and unproctored exams were administered could influence the exam scores on unproctored exams more than on proctored exams. For further analysis, I used the scores of the main group specifically because all students in this group took all four exams with all security mechanisms utilized by the department. To test whether the order effect was significant, I utilized a two-way repeated measures ANOVA with the order in which the exams were administered, first versus second, as the first within-subjects factor and format, proctored versus unproctored, as the second within-subjects factor. To apply the statistical technique, I verified whether the assumptions of the two-way repeated measures ANOVA were met.

The validity of the results of a two-way repeated measures ANOVA depends on the following assumptions: (a) there is no dependency in the scores between participants, (b) the dependent variable score is normally distributed for each level of within-subject factor, (c) the population variance of difference scores computed between any two levels of a within-subject factor is the same (sphericity). The sphericity assumption is not applicable if there are only two levels of the independent variable, as it is in RQ2. ANOVA is robust to violations of the normality assumption if the sample size is large (Field, 2013). In the given study, there was no dependency in the scores between the participants on all exams. To determine whether the dependent variable score was normally distributed across all exams, I ran graphical and numerical analysis of normality in SPSS for the four exams' scores of the students in the main group.

According to the boxplots, the distributions of the scores were slightly left-skewed on all four exams. There was one mild outlier of 4% on S2UPE; all other distributions did not have outliers. Shapiro-Wilk test, a normality test in SPSS, was significant ($p \leq .004$) in all four distributions. The skewness and kurtosis for all four distributions were in acceptable range (S1PE: skewness = -0.52, kurtosis = -0.55; S1UPE: skewness = -0.66, kurtosis =-0.20; S2UPE: skewness = -0.33, kurtosis =-0.52; S2PE: skewness = -0.34, kurtosis =-0.62) and could not affect the validity of the used statistical test. Therefore, the two-way repeated measures ANOVA assumptions were not violated. I ran the test in SPSS.

Results of Testing the Null Hypothesis of RQ2

A two-way repeated measures ANOVA was conducted to test the order effect. The first independent variable order involved in this test had two levels: administered first versus administered second. The second independent variable format also had two levels: proctored versus unproctored. The exam score was the dependent variable. According to the ANOVA results, there was a small significant order effect ($F(1,524) = 45.26$, $p \leq .004$, $\eta^2 = .08$). Thus, the null hypothesis of RQ2 was rejected. I also observed a medium significant format effect ($F(1,524) =66.70$, $p \leq .004$, $\eta^2 = .11$) and a medium significant order*format interaction effect ($F(1,524) = 70.40$, $p \leq .004$, $\eta^2 = .13$). The significant format effect in the two-way ANOVA paralleled the results obtained in RQ1analysis, according to which the format effect in Set 2 was significant.

The profile plots supported significant interaction effect between the order and format: the proctored and unproctored lines intersected. Moreover, the line that corresponded to the proctored exams had a small slope, indicating that the scores on the proctored exams were close to each other. The slope of the unproctored exams' line was much bigger, suggesting large change in scores. These results paralleled my observations during the descriptive statistics stage. Next, I tested the course delivery mode effect to answer RQ3.

Testing the Hypothesis of RQ3

The third research question and the corresponding hypotheses were stated in the following form:

When equivalent automatically-scored web-based exams with the same security mechanisms are used,

RQ3: Is there a relationship between the course delivery mode (IV3), (a) web-assisted face-to-face, (b) hybrid, (c) fully online, and students' scores (DV)?

> $H_0$3: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the course delivery mode.
>
> H_A3: There is a significant difference in students' performance on automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the course delivery mode.

The first independent variable involved in RQ3 was exam format, which is a within-subject variable with two levels: proctored versus unproctored. The second independent variable was course delivery mode, which is a between-subject variable with three levels: face-to-face, hybrid, and online. Because the number of students was different on each exam, I conducted RQ3 analyses separately in Set 1 and Set 2. The course delivery mode analysis for the combined set, which included the scores of the students who took all four exams, is described in the Additional Statistical Tests section. In each set, I began with the descriptive statistics analysis of exam scores across the course delivery modes in the entire population of students under investigation (the total group), narrowing the analysis to the scores of the participants who took the exams with all security mechanisms utilized by the department (the main group).

On S1PE, which was administered first, in the total group, the mean score of the face-to-face students (*N* = 692) was 65.42%. On the same test, the students who were enrolled in the hybrid sections (*N* = 91) had the mean score of 71.19%. The online students (*N* = 55) had the mean score of 71.33%. On S1PE, there were 28 students with extended test time, all of whom were enrolled in face-to-face sections. After I removed the scores of these 28 students, in the main group, the mean score of face-to-face students (*N* = 664) increased very slightly from 65.42% to 65.88%. Because on S1PE there were no students with extended test time in the hybrid and online sections, the descriptive statistics of exam scores in these sections after the removal stayed exactly the same. The summary of the descriptive statistics of the scores on S1PE across the course delivery modes is presented in Table 7.

Table 7

Descriptive Statistics of S1PE Scores with respect to the Course Delivery Modes in the Total and Main Groups

Measure	TG f2f	TG hyb	TG onl	MG f2f	MG hyb	MG onl
N	692	91	55	664	91	55
M	65.42	71.19	71.33	65.48	71.19	71.33
SD	21.48	20.43	22.45	21.41	20.43	22.45

Note: TG = total group; MG = main group; *M* = the mean score on S1PE in %.

According to Table 7, the removal of the scores of students with extended test time on S1PE changed the descriptive statistics in face-to-face group by less than 0.5%. The hybrid and online students had very close means. On average, the scores of face-to-face students were about 5.9% lower than the scores of the hybrid and online students.

The SDs were similar across the curse delivery modes and ranged from 20.43% to 22.45%.

On S1UPE, which was administered second, the students in the face-to-face sections (*N* = 667) had the mean of 67.62%, while the examinees in the hybrid sections (*N* = 89), on average, earned 65.96%. The online students (*N* = 51) had the mean score of 74.35%. After I removed the scores of the students who took the second version of the exam and the students who had extended test time, the means across the delivery modes increased very slightly (face-to-face: *M* = 68.51%; hybrid: *M* = 66.51%; online: *M* = 74.68%). The descriptive statistics for S1UPE scores with respect to the course delivery modes in the total and main groups is summarized in Table 8.

Table 8

Descriptive Statistics of S1PEU Scores with respect to the Course Delivery Modes in the Total and Main Groups

Measure	TG f2f	TG hyb	TG onl	MG f2f	MG hyb	MG onl
N	667	89	51	584	86	48
M	67.62	65.96	74.35	68.51	66.51	74.68
SD	21.11	22.56	19.40	20.95	20.20	18.57

Note: TG = total group; MG = main group; *M* = the mean score on S1PE in %.

According to the statistics shown in Table 6, the removal of scores of the students who took the second version of S1UPE and the students with extended test time increased the mean scores by less than 1% in all course delivery modes. Similar to S1PE, the highest mean was in online sections and lowest one was in the face-to-face group. In the total group, the scores of the hybrid students varied the most, and the scores of online students had the smallest SD. In the main group, the SDs in the face-to-face and hybrid sections were similar, and the SD of the online students was the smallest again.

In the total group with paired S1PE and S1UPE scores, the face-to-face students (*N* = 638) and online students (*N* = 87) had slightly higher mean scores on unproctored exams (face-to-face: M_{S1PE} = 67.32%, M_{S1PE} = 68.27%; online: M_{S1PE} = 73.69%, M_{S1PE} = 74.68%), while the students in the hybrid sections (*N* = 51) had higher mean score on the proctored exam (M_{S1PE} = 71.27%, M_{S1PE} = 65.57%). After I removed the scores of the students with extended test time and students who took version 2 of the proctored exam, the means across the course delivery modes slightly increased. However, the relationship between the scores on the proctored and unproctored exams did not change: the mean scores on the unproctored exam were higher than on the proctored exam in face-to-face and online sections, and lower in the hybrid one (f2f: *N* = 575, M_{S1PE} = 67.67%, M_{S1UPE} = 68.87%, hyb: *N* = 85, M_{S1PE} = 71.35.67%, M_{S1UPE} = 66.36%, onl: *N* = 48, M_{S1PE} = 73.59%, M_{S1UPE} = 74.68%). The descriptive statistics of the paired scores in Set 1 with respect to the course delivery mode in the total and main group is provided in Table 9.

Table 9

Descriptive Statistics of Set1 Exam Scores Across the Course Delivery Modes in Total and Main Groups

Group	*N*	M_{UPE}	SD_{UPE}	M_{PE}	SD_{PE}	*Diff*
S1 TG f2f	638	67.32	20.42	68.27	20.85	-.95
S1 TG hyb	87	71.27	20.71	65.56	22.39	5.71
S1 TG onl	51	73.48	21.31	74.35	19.40	-.87
S1 MG f2f	575	67.67	20.56	68.87	20.87	-1.2
S1 MG hyb	85	71.35	20.53	66.36	20.04	4.99
S1 MG onl	48	73.59	20.94	74.68	18.57	-1.09

Note: TG = total group; MG= main group; M_{PE} = the mean score on S1PE in %; M_{UPE} = the mean score on S1PE in %, *Diff* = $M_{PE} – M_{UPE}$.

According to Table 9, the relationship between the format in which the exams were administered and the exam score was different across the course delivery mode. In the total and main groups, the face-to-face and online students had higher scores on the unproctored exam, while the hybrid students performed better on the proctored test. In the main group, the biggest increase in scores from proctored to unproctored exam administrations was 1.2% and occurred in the face-to-face sections, while the biggest decrease of 4.99% took place in the hybrid sections. To test whether the difference in scores on the proctored and unproctored exams with respect to the course delivery mode was significant, I applied a mixed ANOVA to the scores of the main group only. A mixed ANOVA is appropriate to test hypotheses which involve within-subjects and between-subjects variables, as it is in RQ3 hypotheses. I verified a mixed ANOVA assumptions before I ran the test in SPSS.

Because a mixed ANOVA is a combination of a within-subject ANOVA (a repeated measures ANOVA) and between-subject ANOVA, the validity of the results of a mixed ANOVA depends on the following assumptions: (a) there is no dependency in the scores between participants, (b) the dependent variable score is normally distributed for each level of a within-subject factor, (c) The variances of the dependent variable across the levels of a between-subject factor are the same (homogeneity), (d) The population variance of difference scores computed between any two levels of a within-subject factor is the same (sphericity). The sphericity assumption is not applicable if there are only two levels of a within-subject factor, as it is in RQ3. ANOVA is robust to violations of normality if the sample size is at least 30 (Field, 2013).

In the given study, there was no dependency in the scores between the participants on all exams. To determine whether the dependent variable score was normally distributed on Set 1 exams with respect to the course delivery modes, I ran graphical and numerical analysis of normality in SPSS for the paired exams scores in the main group of Set 1. To identify whether homogeneity was met, I used the Levene's test.

According to the box-plots, the distributions of scores on both Set 1 exams were slightly left-skewed in all course delivery modes. Three outliers of 4%, 6%, and 8% were observed on S1UPE in the face-to-face group. Shapiro-Wilk test was not significant in S1UPE hybrid group (*p* =.090, *N* = 85) and S1UPE online group (*p* =.118, *N* = 48), but significant in S1PE face-to-face group (*p*<.001, *N* = 575), S1PE hybrid group (*p* =.003), S1PE online group (*p* =.048), and S1UPE face-to-face group. The skewness and kurtosis for all six distributions were in acceptable range (S1PE f2f: *skewness* = -0.32, *kurtosis* =-0.85; S1PE hyb: *skewness* = -0.57, *kurtosis* =-0.60; S1PE onl: *skewness* = -0.78, *kurtosis* =-0.024 S1UPE f2f: *skewness* = -0.63, *kurtosis* =-0.23; S1UPE hyb: *skewness* = -0.40, *kurtosis* =-0.76; S1UPE onl: *skewness* = -0.40, *kurtosis* =-0.76). The normality assumption was not violated.

According to the Levene's test of homogeneity, the variances across the course delivery modes were not significantly different on both Set1 exams (S1PE: *F*(2,705) =.05, *p* =.951; S1PE: *F*(2,705) =1.12, *p* =.328).The sphericity assumption was not applicable to this test because the independent variable format has only two levels. All mixed ANOVA assumptions were met; I ran the test in SPSS.

Results of Testing the Null Hypothesis of RQ3 in Set 1

According to the ANOVA results, there was no significant difference in scores on proctored and unproctored exams in Set 1 across all delivery modes (*F*(2,705) = 2.058, *p*

=.128, $\eta^2 = .006$). Therefore, the course delivery mode effect in Set 1 was not significant. The null hypothesis of RQ3 in Set 1 was retained. The format effect was also not significant ($F(2,705) = 0.903$, $p = .342$, $\eta^2 = .001$), which paralleled the results of RQ1 analysis. However, the interaction format*mode effect was significant with a small effect size ($F(2,705) = 6.034$, $p = .003$, $\eta^2 = .017$). In Set 1, the number of students across the course delivery modes was not the same ($N_{f2f} = 575$, $N_{hyb} = 85$, $N_{onl} = 48$). Unequal sizes of between-subject groups may influence the test results and artificially decrease *p*-values (Field, 2013). For this reason, some adjustments for significant *p*-values close to .05 might be needed (Field, 2013). However, the *p* of .003 was much smaller than .05. Therefore, unequal size in the groups did not influence the found significance.

The significant format*mode interaction suggested that there was a significant difference in how the change in scores on proctored and proctored exams manifested across the course delivery modes. According to SPSS profile plots for the format*mode interaction, in Set 1, the change in scores had similar pattern in the face-to-face and online groups, but different in the hybrid group. This fact corresponds to the previously discussed descriptive statistics analysis, according to which the mean scores in the face-to-face and online groups were higher on the unproctored exam than on proctored one, while in the hybrid group the students performed better on the proctored exam (f2f: M_P-M_U = -1.20%; hyb: M_P-M_U = 4.99%; onl: M_P-M_U = -1.09%). Interpretations of these results are given in Chapter 5. Next, I conducted the same analyses with the scores in Set 2.

In Set 2, 739 students took the first exam, which was unproctored. In this total group, 612 were enrolled in face-to-face sections and had the mean score of 62.10% on S2UPE. The mean scores of the hybrid students (*N* = 81) and online students (*N* = 46) were almost the same ($M_{hy\ b}$ = 64.60%, M_{onl} = 64.40%). On S2UPE, 82 students took the second version of the exam and 20 had extended test time. After I removed the scores of these 82 and 20 students, the mean score of the remaining 513 face-to-face students, 78 hybrid students, and 46 online students became 62.77%, 64.87%, and 64.40% respectively. The summary of the descriptive statistics of S2UPE in the total and main group is provided in Table 10.

Table 10

Descriptive Statistics of S2UPE Scores with respect to the Course Delivery Modes in the Total and Main Groups

Measure	TG f2f	TG hyb	TG onl	MG f2f	MG hyb	MG onl
N	612	81	46	513	78	46
M	61.60	64.57	64.40	62.77	64.87	64.40
SD	19.81	19.45	25.00	19.64	19.33	25.00

Note: TG = total group; MG = main group; *M* = the mean score on S2UPE in %.

As seen in Table 10, after the removal of the scores of the students who took the second version of the exam and the students who had extended test time, the statistics in the online sections stayed the same because the number of students in this group did not change. The mean in the face-to-face sections increased by 1.2% and in the hybrid sections by 0.3%. The SDs were similar in the face-to-face and hybrid sections, but higher in the online sections in both groups. Thus, the total and main groups had similar descriptive statistics on S2UPE with respect to the course delivery modes.

In Set 2, the same number of students, 739, took the second exam. Out of these 739 students, 611 were face-to-face students, whose mean score on S2PE was 68.25% (*SD* = 17.86%), 82 were hybrid students with the mean score of 69.71% (*SD* = 20.53%) and 46 were online students whose mean score was 69.56% (SD = 20.42%). This second exam in Set 2 had to be proctored. However, 55 of 739 students took the exam in unproctored format. After I removed the scores of these 55 studemts, all of whom were enrolled in face-to-face section, 684 students remained. Out of these 684 students, 556 were enrolled in face-to-face sections and had the mean score of 68.17% (*SD* = 17.59). The removal of the 56 students reduce the means score of fase-to-face students by .08%. The additional removal of 13 students with extended test time, all of whom were also enrolled in face-to-face sections, increased the mean score of face-to-face students (*N* = 543) back to the initial 68.25% (SD = 17.54) of the total group. The decriptve statistics of the scores on S2PE in the total and main groups with respect to the course delivery modes is presented in Table 11.

Table 11

Descriptive Statistics of S2PE Scores with respect to the Course Delivery Mode in the Total and Main Groups

Measure	TG f2f	TG hyb	TG onl	MG f2f	MG hyb	MG onl
N	611	82	46	543	82	46
M	68.25	69.71	68.56	68.25	69.71	68.56
SD	17.86	20.53	20.42	17.54	20.53	20.42

Note: TG = total group; MG = main group; *M* = the mean score on S2PE in %.

As seen in Table 11, on S2PE, the descriptive statistics in the total and main groups were identical. With respect to the delivery mode, the mean scores were very similar: the biggest score in the hybrid sections just 1.46% higher than the smallest mean in the face-to-face sections. The SDs were very close in the hybrid and online sections, while the SD in the face-to-face sections was about 3% lower.

In Set 2, 725 students took both exams. Out of these 725 in the total group, 598 were face-to-face students, 81 were enrolled in the hybrid sections, and 46 were online students. After I removed the scores of the students who took both exams in Set 2 in unproctored format, students who completed the second version of S2UPE, and students with the extended test time, the remaining main group included 593 students. Out of these 577 students, 455 were enrolled in face-to-face classes, 78 were hybrid students, and 46 were online students. The descriptive statistics of the paired scores in Set 2 in the total and main group is shown in Table 12.

Table 12

Descriptive Statistics of Set2 Exam Scores Across the Course Delivery Modes in Total and Main Groups

Group	*N*	M_{S2UPE}	SD_{S2UPE}	M_{S2PE}	SD_{S2PE}	*Diff*
S2 TG f2f	598	62.18	19.50	68.79	17.44	6.61
S2 TG hyb	81	64.57	19.45	69.87	20.61	5.30
S2 TG onl	46	64.40	25.00	69.56	20.42	5.16
S2 MG f2f	469	63.59	19.43	69.46	17.05	5.87
S2 MG hyb	78	64.87	19.33	69.80	20.53	4.93
S2 MG onl	46	64.40	25.00	69.56	20.42	5.16

Note: TG = total group; MG= main group; M_{S2PE} = the mean score on S2PE in %; M_{S2UPE} = the mean score on S2PE in %, *Diff* = $M_{S2PE} – M_{S2UPE}$.

According to Table 12, in Set 2, the mean scores in the total group and main group were almost the same in hybrid and online sections, but the means in the total group were lower than the means in the main group in face-to-face classes. In both groups the mean scores on the proctored exam, which in Set 2 was administered second, were higher than the means on the unproctored exam, with approximtely the same increase of 5%-6%. On both exams, hybrid and online students had almost the same mean scores, while the mean scores of face-to-face students were lower. On S2UPE, the SDs were almost the same in the face-to-face and hybrid sections, but about 5.8% bigger in the online sections. In contrast, on S2PE, the SDs were close in hybrid and online classes, but about 3.5% lower in face-to-face sections. Next, I conducted another mixed ANOVA for the paired scores in the main group. The assumptions of mixed ANOVA in Set 2 were verified before I ran the test.

According to the box-plots, the distributions of scores on both Set 2 exams were slightly left-skewed in all course delivery modes. There was one outlier of 4% in S2UPE face-to-face group and one outlier of 10% in S2PE hybrid group. Shapiro-Wilk test was significant in all groups (S2UPE f2f: *p* <.001; S2UPE hyb: *p* =.048; S2UPEonl: *p* =.004; S2PE f2f: *p* <.001; S2PE hyb: *p* =.002; S2PEonl: *p* =.022). The skewness and kurtosis for all six distributions were in acceptable range (S1PE f2f: *skewness* = -0.26, *kurtosis* =-0.59; S1PE hyb: *skewness* = -0.74, *kurtosis* =-0.07; S1PE onl: *skewness* = -0.41, *kurtosis* =-0.97 S1UPE f2f: *skewness* = -0.31, *kurtosis* =-0.42; S1UPE hyb: *skewness* = -0.50, *kurtosis* =-0.27; S1UPE onl: *skewness* = -0.13, *kurtosis* =-1.41) and could not affect the validity of the ANOVA. The normality assumption was met.

Levene’s test of homogeneity was significant on both exams in Set 2 (S2PE: *F*(2,590) = 5.66, *p* =.004; S2PE: *F*(2,590)=4.88, *p* =.008), suggesting that the variances across the course delivery modes were significantly different on S2PE and on S2UPE. However, if the sizes of groups are large, Levene’s test can be significant when the variances are practically equal (Field, 2013). According to Field (2013), if the ratio of the largest SD to the smallest SD is less than two for each level of the independent variable, the homogeneity assumption is not violated. In Set 2, there were 469 face-to-face students, 78 hybrid students, and 46 online students. On S2PE, the largest SD of 25.00 was in the online group and the smallest SD of 19.33 was in the hybrid group. On S2UPE, the largest SD of 20.42 was in the hybrid group, and the smallest SD of 17.05 was in the face-to-face group. The ratio of 25.00 to 19.33 is 1.3; the ratio of 20.42 to 17.05 is 1.2. Therefore, the variances across the groups in Set 2 were not significantly different and all mixed ANOVA assumptions were met. I ran the test in SPSS.

Results of Testing the Null Hypothesis of RQ3 in Set 2

In Set 2, the format effect was significant ($F(1,592) = 36.62$, $p < .001$, $\eta^2 = .058$), which paralleled the findings of RQ1 analysis in Set 2. The course delivery mode effect ($F(2,590) = .079$, $p = .924$, $\eta^2 < .001$) was not significant. The null hypothesis of RQ3 in Set 2 was retained. The interaction format*mode effect $F(2,590) = .189$, $p = .828$, $\eta^2 = .001$) was also not significant. The profile plots for the format*mode interaction, in Set 2, had similar pattern for the change in scores across the course delivery modes: all three lines were almost parallel to each other. These findings corresponded to the descriptive statistics analysis, according to which the students had different mean scores on S2PUE and S2PE in all course delivery modes, and the scores increased from the first to the second test administrations by about 5.5% in all groups. To compare the findings of RQ3 analyses in Set 1 and Set 2, I combined the mixed ANOVA results for both sets in Table 13.

Table 13

Results of RQ3 Analysis in Set 1 and Set 2

Effect	*N*	*F*	*p*	η^2
S1 mode	708	2.06	.13	.01
S1 format		.90	.34	.00
S1 format*mode		6.03	.00	.02
S2 mode	593	.08	.92	.00
S2 format		36.62	.00	.06
S2 format*mode		.19	.83	.00

Note: S1 = Set1; S2 = Set 2.

As shown in Table 13, the mode effect was not significant in both sets: the null hypothesis of RQ3 was rejected in Set 1 and Set 2. The format effect was not significant in Set 1, but significant in Set 2: the null hypothesis of RQ1 was retained in Set 1, but rejected in Set 2. The format*mode interaction effect was significant in Set 1, but not significant in Set 2. Interpretations of these findings are given in Chapter 5. I proceeded to RQ4 analyses.

Testing the Null Hypothesis of RQ4

The fourth research question and the corresponding hypotheses were stated in the following form:

When equivalent automatically-scored web-based exams with the same security mechanisms are used,

RQ4: Is there a relationship between the instructor (IV4) and students' scores (DV)?

> $H_0$4: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the instructor of the course.

H_A4: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the instructor of the course.

The within-subject variable involved in RQ4 was exam format with two levels: proctored versus unproctored. The between-subject variable in RQ4 was instructor with seven levels: seven instructors were involved in the study. Due to the fact that the number of students was different on each exam, I performed RQ4 analyses separately in Set 1 and Set 2. The instructor effect analysis in the combined set, which included the scores of the students who took all four exams, is given in the Additional Statistical Tests section. Similar to all previous analyses, in each set, I began with the discussion of the descriptive statistics of exam scores across the instructors in the entire population of students under investigation (the total group), narrowing the analysis to the scores of the participants who took the exams with all security mechanisms utilized by the department (the main group).

In the total group, 838 students took S1PE. After I removed the scores of 28 students with extended test time from the total group, in the resulting main group, the number of the students of Instructor 1 stayed the same, while the number of the students of all other instructors slightly decreased. The descriptive statistics of the scores on S1PE with respect to the instructors in the total and main groups is provided in Table 14.

Table 14

Descriptive Statistics of S1PE Scores with Respect to Instructors in Total and Main Groups

Groups	I_1	I_2	I_3	I_4	I_5	I_6	I_7
TG *N*	20	109	327	80	149	76	77
TG *M*	54.87	67.63	69.01	57.78	64.62	69.35	66.43
TG *SD*	24.23	20.63	20.62	23.24	19.75	23.44	22.42
MG *N*	20	103	321	71	145	75	75
MG *M*	54.87	68.08	69.06	58.27	65.05	69.60	67.66
MG *SD*	24.23	20.54	20.71	23.77	19.65	23.50	21.35

Note: TG = the total group; MG = the main group; *M* = mean score on S1PE in %, I = instructor

According to Table 14, the removal of the scores of the 28 students increased the mean scores of the students of all instructors, except Instructor 1, whose number of students stayed the same. The change in the mean scores ranged from .05% to 1.23%. On S1PE, the students performed differently across the instructors with the difference of about 14.7% between the lowest and highest mean.

In Set 1, 807 students took S1UPE. After I excluded the students who took the second version of the exam and the students with extended test time, the number of students decreased in all groups with respect to the instructors, except Instructor 1. The descriptive statistics of the scores on S1UPE with respect to the instructors in the total and main groups is presented in Table 15.

Table 15

Descriptive Statistics of S1UPE Scores with Respect to Instructors in Total and Main Groups

Groups	I_1	I_2	I_3	I_4	I_5	I_6	I_7
TG *N*	19	103	319	74	145	74	73
TG *M*	62.44	69.05	67.86	62.66	66.26	70.07	73.13
TG *SD*	24.32	18.68	21.41	24.53	19.96	23.02	19.71
MG *N*	19	88	298	56	123	66	68
MG *M*	62.44	69.46	68.58	64.32	67.93	70.29	73.54
MG *SD*	24.32	18.95	20.94	24.51	19.18	23.08	20.33

Note: TG = the total group; MG = the main group; *M* = mean score on S1UPE in %, I = instructor

As seen in Table 15, after the exclusion, the mean scores of the students of all instructors, except Instructor 1, increased. The change ranged from 0.22% to 1.67%. On S1UPE, on average, the participants performed differently across the instructors with the difference of 11.1% between the lowest and highest mean.

There were 795 students who took both exams in Set 2. I removed the scores of 52 students who took version 2 of S1UP exam and 11 students with extended time. The resulting main group had 732 students. The descriptive statistics of the paired scores in Set 1 with respect to the instructors in the total and main groups is shown in Table 16.

Table 16

Descriptive Statistics of Set 1 Scores with Respect to Instructors in the Total and Main Groups

Groups	I_1	I_2	I_3	I_4	I_5	I_6	I_7
TG *N*	19	104	317	73	140	70	72
TG M_{S1PE}	57.13	67.35	69.59	57.30	64.37	72.60	66.58
TG SD_{S1PE}	22.63	20.51	20.31	22.21	19.56	20.54	21.80
TG M_{S1UPE}	65.18	68.83	68.08	62.80	65.94	71.99	72.23
TG SD_{S1UPE}	23.62	18.72	21.24	24.57	20.09	21.79	20.27
MG *N*	19	92	304	66	121	62	68
MG M_{S1PE}	57.13	66.56	69.74	57.59	65.18	73.25	67.57
MG SD_{S1PE}	22.63	20.92	20.23	23.93	19.42	21.02	21.87
MG M_{S1UPE}	65.18	69.46	68.41	63.54	67.32	72.48	72.42
MG SD_{S1UPE}	23.62	18.73	21.07	25.12	19.84	21.65	20.82
Diff *N*	0	12	14	6	19	8	4
Diff M_{S1PE}	0	.79	-.15	-.29	-.81	-.65	-.99
Diff M_{S1UPE}	0	-.63	-.33	-.74	-1.38	-.49	-.19

Note: TG = the total group; MG = the main group; M_{S1PE} = mean score on S1PE in %, M_{S1UPE} = mean score on S1UPE in %; Diff N = TG *N* - MG *N*; Diff M_{S1PE} = TG M_{S1PE} - MG M_{S1PE}; Diff M_{S1UPE} = TG M_{S1UPE}- MG M_{S1UPE}; I = instructor

According to Table 16, the exclusion of the scores of the students who took version 2 of S1UPE and students who had extended test time changed the mean scores with respect to instructors on the proctored exam by less than 1% and on the unproctored exams by less than 1.5%. For further analyses, I used the main group only, in which the students of all instructors had higher scores on S1UPE than on S1PE, except Instructor 3 and Instructor 6’s students. The biggest increase in scores from S1PE to S2UPE was in Instructor 1 and Instructor 7’s groups (*diff* I_1= 8.05%, *diff* I_2 = 2.9%, *diff* I_3 =-1.33%, *diff* I_4 =5.95%, *diff* I_5 =2.14%, *diff* I_6 = -.77 %, *diff* I_7 = 6.85%). To test wether the observed differences in the mean scores with respect to the instructors were significant, I conducted another mixed ANOVA, the assumptions of which were tested before I ran the test.

According to the boxplots, all 7 distributions of the scores on S1PU with respect to the instructors were slightly skewed to the left without outliers. Shapiro-Wilk test for Set 1 proctored exam was not significant in Instructor 1 group (*p* =.200, *N* =19), Instructor 2 group (*p* =.200), and Instructor 6 group (*p* =.052), but was significant in all others (I3: *p* <.001; I4: *p* =.022; I5: *p* =.046; I7: *p* <.001). The skewness and kurtosis of all seven distributions were in acceptable limits (I1: *skewness* = -0.25, *kurtosis* =-1.29; I2: *skewness* = -0.31, *kurtosis* =-0.63; I3: *skewness* = -0.58, *kurtosis* =-0.30; I4: *skewness* = -0.86, *kurtosis* =-0.15; I5: *skewness* = -0.38, *kurtosis* =-0.43; I6: *skewness* = -0.92, *kurtosis* =-0.05; I7: *skewness* = -0.75, *kurtosis* =-0.07).

The box plots for the distributions of S1UPE for all seven instructors were also left-skewed. There were two mild outliers of 6% and 8% in Instructor 3 group and one mild outliers of 20% in Instructor 6 group. All other distributions did not have outliers. On S1UPE, Shapiro-Wilk test was not significant in Instructor 1 group (*p* =.09), but significant in all other distributions (I2: *p* =.002; I3: *p* < .001; I4: *p* <.001; I5: *p* =.023; I6: *p* <.001; I7: *p* =.001). The skewness and kurtosis of all seven distributions were in acceptable limits (I1: *skewness* = -0.27, *kurtosis* =-1.43; I2: *skewness* = -0.30, *kurtosis* =-1.05; I3: *skewness* = -0.58, *kurtosis* =-0.30; I4: *skewness* = -0.86, *kurtosis* =-0.15; I5: *skewness* = -0.38, *kurtosis* =-0.43; I6: *skewness* = -0.92, *kurtosis* =-0.05; I7: *skewness* = -0.75, *kurtosis* =-0.07). The normality assumption of the scores on both Set 1 exams across all instructors was not violated.

Levene’s test of homogeneity was not significant on S1UPE (*F*(6,725) =1.57, *p* =.15), but significant on S1PE (*F*(6,725) = 2.34, *p* =.03). I calculated the ratio of the largest SD to the smallest SD in S1PE and found that the homogeneity assumption was not violated (24.09/19.43 = 1.2). Therefore, all mixed ANOVA assumptions were met. I ran the test for Set 1 exams in SPSS.

Results of Testing the Null Hypothesis of RQ4 in Set 1

In Set 1, there was a small significant format effect (*F*(1, 732) = 7.653, *p* =.006, η^2 = .010), indicating that if all other variables are ignored, scores on proctored and unproctored exams were significantly different. This result contradicted to RQ1, RQ2, and RQ3 findings, according to which there was no significant format effect in Set 1 when the variable instructor was not present. Thus, the addition of the variable instructor changed the format effect from insignificant to significant. There was also a small significant instructor effect (*F*(6, 725) = 3.747, *p* =.001, η^2 = .030): the null hypothesis of RQ4 in Set 1 was rejected. The differences between S1PE and S1UPE scores were significantly different across the instructors. To identify where this difference could occur, I conducted a pairwise comparison test in SPSS. According to the pairwise comparison, the differences in scores on S1PE and S1UPE were not significant in all pairs of instructors except Instructor 4 and Instructor 3 ($Diff_{I4\text{-}I3}$ = *-8.513%*, *p* =.008), Instructor 4 and Instructor 6 ($Diff_{I4\text{-}I6}$ =*-12.302%*, *p* =.002), and Instructor 4 and Instructor 7 ($Diff_{I4\text{-}I7}$ = *-9.430%*, *p* =.043). Interpretations of these results are given in Chapter 5.

However, there was no significant interaction between the format in which exams were administered and instructors (*F*(6, 725) = 1.90, *p* =.079, η^2 = .015), suggesting that the change in scores from proctored to unproctored exams did not manifest significantly different across the instructors. This fact was supported by the corresponding profile plot, according to which, the change in scores had similar increasing pattern in scores from proctored to unproctored exam in all groups with respect to instructors, except Instructor 3 and Instructor 6, the students of whom had slightly higher scores on proctored exams. Next, I conducted the same analyses in Set 2.

In Set 2, in the total group, 739 students took the unproctored exam, which was administered first. After I excluded the students who took the second version of the exam and the students who had extended test time, there were 637 in the main group. The descriptive statistics of S2UP scores with respect to instructors in the total and main groups is provided in Table 17.

Table 17

Descriptive Statistics of S2UPE Scores with Respect to Instructors in Total and Main Groups

Groups	I_1	I_2	I_3	I_4	I_5	I_6	I_7
TG *N*	14	97	295	61	136	68	68
TG *M*	65.89	63.93	64.35	63.17	58.25	58.95	58.84
TG *SD*	17.88	17.93	20.25	21.66	19.96	21.91	18.83
MG *N*	13	78	282	55	88	60	61
MG *M*	65.50	64.59	64.62	63.92	60.63	61.10	58.53
MG *SD*	18.46	17.77	20.36	21.89	19.71	20.82	18.81
Diff	.39	-.66	-.27	-.75	-2.38	-2.15	.31

Note: TG = the total group; MG = the main group; *M* = mean score on S2UPE in %, I = instructor

According to Table 17, S2UPE scores in the total and main groups did not differ by more than 2.4%. The difference between the largest and smallest SDs was about 4% in both groups. Thus, the exclusion did not change the descriptive statistics of S2UPE a lot.

The second exam in Set 2 was taken by the same number of students as the first one (*N* = 739). After I removed the students with extended test time and students who took both exams in Set 2 in unproctored format, there were 671 students left. The descriptive statistics of S2PE scores with respect to the instructors in the total and main groups is presented in Table 18.

Table 18

Descriptive Statistics of S2PE Scores with Respect to Instructors in Total and Main Groups

Groups	I1	I2	I3	I4	I5	I6	I7
TG *N*	13	93	296	64	137	66	70
TG *M*	68.54	66.12	69.23	67.52	67.90	69.74	69.34
TG *SD*	16.69	17.25	19.03	18.26	17.20	18.23	19.49
MG *N*	13	92	290	62	136	64	14
MG *M*	68.54	65.95	69.27	67.85	67.82	70.39	70.91
MG *SD*	16.69	17.77	19.06	17.91	17.24	18.07	15.47
Diff	0	.17	-.04	-.33	-.08	-.65	-1.57

Note: TG = the total group; MG = the main group; *M* = mean score on S2PE in %; Diff = TG *M* - MG *M* ; I = instructor

According to Table 18, the scores on the second exam in Set 2 in the total and main groups were very similar. The difference between the largest and smallest SD on S2PE was about 2% in both groups, lower than on S2UPE. Thus, similar to S2UPE, the removal of the scores changed the descriptive statistics of S2PE very slightly. Next, I proceeded to the analysis of the paired scores in Set 2.

The total group of the paired scores in Set 2 included the scores of 725 students. After all exclusion, there were the scores of 593 students left. The descriptive statistics of the paired scores in Set 2 with respect to instructors in the total and main group is shown in Table 19.

Table 19

Descriptive Statistics of Set 2 Scores with Respect to Instructors in the Total and Main Groups

Groups	I_1	I_2	I_3	I_4	I_5	I_6	I_7
TG *N*	13	93	293	60	134	64	68
TG M_{S2PE}	68.54	66.12	69.50	69.71	68.25	70.49	69.78
TG SD_{S2PE}	16.69	17.25	18.91	16.23	16.88	17.44	19.56
TG M_{S2UPE}	66.85	65.51	64.38	63.89	58.53	60.56	58.84
TG SD_{SUPE}	18.23	16.39	20.31	21.10	19.80	21.49	18.83
MG *N*	12	80	286	58	88	58	11
MG M_{S2PE}	68.35	66.21	69.77	70.00	70.12	71.17	71.97
MG SD_{S2PE}	17.41	17.03	18.67	16.14	17.29	17.39	15.64
MG M_{S2UPE}	67.58	65.53	64.65	64.51	60.63	62.02	57.09
MG SD_{S2UPE}	18.84	17.03	20.32	21.04	19.71	20.56	19.92
Diff *N*	1	13	7	2	46	6	57
Diff M_{S2PE}	.19	-.09	-.27	-.29	-1.87	-.68	-2.19
Diff M_{S2UPE}	-.73	-.02	-.27	-.62	-2.10	-1.46	1.75

Note: TG = the total group; MG = the main group; M_{S2PE} = mean score on S2PE in %, M_{S2UPE} = mean score on S2UPE in %; Diff N = TG *N* - MG *N*; Diff M_{S2PE} = TG M_{S2PE} - MG M_{S2PE}; Diff M_{S2UPE} = TG M_{S2UPE}- MG M_{S2UPE}; I = instructor

According to Table 19, the exclusion of the scores of the students who took version 2 of S2UPE, students who had extended test time, and students who took both exams in Set 2 in unproctored format increased the mean scores for all instructors except Instructor 1 on S2PE and Instructor 7 on S1UPE.The biggest change in the mean scores of 2.2% and 2.1% was in Instructors 7 and Instructor's 5 groups respectively. In the majority of the groups, the change was less than 1%.

For further analysis, I used the paired scores in the main group only (*N* = 593). In this group, the students of all instructors had higher scores on S2PE than on S2UPE. The biggest increase in scores from S2UPE to S2PE was in Instructor 5, Instructor 6, and Instructor 7's groups (*diff* I_1= -0.77%, *diff* I_2= -0.68%, *diff* I_3=-5.12%, *diff* I_4=-5.49%, *diff* I_5=-9.49%, *diff* I_6 = - 9.15%, *diff* I_7= -14.88 %). To test whether the observed differences in the mean scores with respect to the instructors in Set 2 were significant, I conducted one more mixed ANOVA. The assumptions were verified before I ran the test.

According to the boxplots, on S2UPE, the distributions of students' scores of Instructor 1, Instructor 3, and Instructor 7 were almost symmetric. However, there was one outlier of 13.5% in Instructor 7 group. All other distributions of the scores on S2UPE were slightly left-skewed without outliers. On S2PE, the distributions of students' scores of Instructor 2, Instructor 4, and Instructor 5 were almost symmetric; the distributions of the scores on S2PE in all other groups were slightly left-skewed. There was one mild outlier of 10% in Instructor 3 group.

On S2UPE, Shapiro-Wilk normality test was not significant in Instructor 1 (p = .64), Instructor 2 (p = .49), Instructor 5 (p =.52), Instructor 6 (p =.16), and Instructor 7 (p = .82) groups, but significant in Instructor 3 ($p \leq$.004) and Instructor 4 (p =.03) groups. On S2PE, Shapiro-Wilk test was not significant in Instructor 1 (p = .40), Instructor 2 (p = .56), Instructor 4 (p =.26), Instructor 5 (p = .15), and Instructor 7 (p = .74) groups, but significant in Instructor 3 (p <.001) and Instructor 6 (p = .02) groups.

On S2UPE, the skewness and kurtosis of all seven distributions were in acceptable limits (I1: *skewness* = -0.52, *kurtosis* =-0. 63; I2: *skewness* = -0.001, *kurtosis* =-0.70; I3: *skewness* = -0.33, *kurtosis* =-0.63; I4: *skewness* = -0.18, *kurtosis* =-1.10; I5: *skewness* = -0.27, *kurtosis* =-0.30; I6: *skewness* = -0.55, *kurtosis* =-0.02; I7: *skewness* = -0.81, *kurtosis* =-1.32). Similarly, on S2PE, skewness and kurtosis had acceptable range (I1: *skewness* = -0.08, *kurtosis* =-1.46; I2: *skewness* = -0.02, *kurtosis* =-0.41; I3: *skewness* = -0.55, *kurtosis* =-0.33; I4: *skewness* = -0.05, *kurtosis* =-0.88; I5: *skewness* = -0.26, *kurtosis* =-0.39; I6: *skewness* = -0.50, *kurtosis* =-0.71; I7: *skewness* = -0.27, *kurtosis* =-0.74). The normality assumption was not violated. There were two groups with small sizes: Instructor 1 (*N* =12) and Instructor 7 (*N* = 11). However, I kept them in the analysis because their distributions were almost symmetric and did not have significant Shapiro-Wilk test on both exams in Set 2.

In Set 2, Levene's test of homogeneity was not significant on both exams (S2UPE: *F*(6,586) =0 .87, *p* =.52; S2PE: *F*(6,586) =0.67, *p* =.68), indicating that the variances across instructors were not significantly different. Thus, all mixed ANOVA assumptions were met. I conducted the test in SPSS.

Results of Testing the Null Hypothesis of RQ4 in Set 2

In Set 2, the instructor effect was not significant *F*(6,586) = .20, *p* =.98, η^2 = .00): the null hypothesis of RQ4 in Set 2 was retained. The format effect (*F*(1,592) = 48.25, *p* $\leq$.004, η^2 = .08) and format*instructor interaction effect *F*(6,587) = 4.83, *p* <.001, η^2 = .05) were significant. Thus, the increase in scores from S2UPE to S2PE observed in the descriptive statistics analysis was significant (significant format effect) and this increase occurred in all groups with respect to instructors (insignificant instructor effect). However, the change in scores manifested significantly differently across instructors (significant format*instructor interaction effect): the increase in scores of the students of Instructor 1 and Instructor 2 were less than 1%, while the increase in scores of other instructors ranged from 5.12% to 14.88%. The corresponding profile plot in SPSS supported the significant format*instructor interaction: the line segments for Instructor 1 and Instructor 2 were parallel, but intersected the line segments of all other instructors.

To compare the findings of RQ4 analyses in Set 1 and Set 2, I combined the mixed ANOVA results for both sets in Table 20.

Table 20

Results of RQ4 Analysis for Set 1 and Set 2

Effect	*N*	*F*	*p*	η^2
S1 instructor	732	3.75	.00	.03
S1 format		7.65	.01	.01
S1 format*instructor		1.90	.08	.02
S2 instructor	593	0.20	.98	.00
S2 format		48.25	.00	.08
S2 format*instructor		4.83	.00	.00

Note: S1 = Set1; S2 = Set 2.

As seen in Table 20, a small significant instructor effect took place on Set 1 exams, but no instructor effect was observed on Set 2 exams. The null hypothesis of RQ4 was rejected in Set 1, but retained in Set 2. The format effect was significant in both sets; however, in Set 1, the effect size was smaller. The format*instructor interaction effect was not significant in Set 1, but significant in Set 2 with a small effect size. Interpretations of these findings are given in Chapter 5.

Additional Statistical Tests

To examine relationships between the variables order and the variables course delivery mode and instructor, I conducted a full mixed ANOVA with all four independent variables involved in the study. Although these interactions were not related to the four research questions and the corresponding hypotheses directly, the results of the full mixed ANOVA allowed for better understanding of RQ1, RQ2, RQ3, RQ4 findings and their interpretations. I completed the descriptive statistics analysis and verified all needed assumptions before I ran the test.

There were 659 students who completed all four exams (total group), but only 525 students took all four exams with all security mechanisms utilized by the department (main group). I conducted all analyses related to the full mixed ANOVA with the main group only. The descriptive statistics of the scores on all four exams across the course delivery modes and instructors is presented in Table 21.

Table 21

Descriptive Statistics of Scores on Four Exams with Respect to Course Delivery Mode and Instructors

Mode	Inst	*N*	$M_{S1PE}(SD)$	$M_{S1UPE}(SD)$	$M_{S2UPE}(SD)$	$M_{S2PE}(SD)$
f2f	I_1	12	70.13 (14.19)	72.93 (20.89)	67.58 (18.84)	68.35 (17.41)
	I_2	68	71.85 (18.76)	73.62 (16.59)	67.03 (16.44)	67.65 (17.06)
	I_3	152	70.11 (18.56)	69.97 (18.91)	64.58 (19.45)	69.83 (17.33)
	I_4	43	66.30 (21.68)	71.79 (19.19)	65.07 (22.23)	70.78 (16.37)
	I_5	72	69.61 (18.56)	70.46 (17.76)	62.11 (18.80)	72.48 (16.57)
	I_6	50	76.52 (20.37)	76.24 (19.51)	61.74 (21.33)	71.70 (17.61)
	I_7	10	78.59 (14.65)	84.46 (10.64)	55.79 (20.51)	69.77 (14.57)
hyb	I_3	75	73.29 (19.40)	68.37 (22.05)	64.99 (19.71)	70.34 (19.59)
onl	I_3	43	76.91 (18.76)	75.99 (18.95)	64.12 (24.91)	69.95 (19.60)

Note: M_{S1PE} = mean score on S1PE in %, M_{S1UPE} = mean score on S1UPE in %; M_{S2PE} = mean score on S2PE in %, M_{S2UPE} = mean score on S2UPE in %; Inst = instructor

According to Table 21, in face-to-face group, when the unproctored exam was administered second, the scores on the unproctored exam were either almost the same or higher than the scores on the proctored exam for all instructors. But when the unproctored exam was administered first, the face-to-face students of all instructors had lower scores on the unproctored test. The biggest difference of 28.67% between the unproctored exam mean score in Set 1 and unproctored exam mean score in Set 2 occurred in Instructor 7 group. The smallest difference in scores of 5.35% on the same exams took place in Instructor 1 group. The hybrid students had higher scores on the proctored exams in both sets with the difference between the scores on proctored and unproctored exams of about 4% in both sets. The difference between the scores on the unproctored exams in this group was about 3.4%. Similarly, the online group, had higher scores on the proctored exams than on the unproctored tests in both sets, but the difference between the scores on the unproctored exams was 11.9%.

According to the boxplots, on all four exams the distributions of scores in Instructor 1 and Instructor 7 groups were almost symmetric, without outliers. The distributions of scores of all other instructors were slightly left-skewed, without outliers except Instructor 6 group, which had a mild outlier of 18% on S1PE and another mild outlier of 17.5 on S1UPE. On S1PE, Shapiro-Wilk test was not significant in Instructor 1, Instructor 5, and Instructor 7 groups on all exams (I1: p =.51; I5: p =.11; I7:p =.30), Instructor 2 group on Set 2 exams (S2PE: p =.49; S2UPE p = .31) and Instructor 4 group on S1UPE (p = .07). For all other distributions of the scores with respect to instructors, Shapiro-Wilk test was significant (all $p \leq .004$). Skewness and kurtosis ranged from -1.2 to 0.07 in all distributions of scores across instructors. The Instructor 1 (N = 12) and Instructor 7 (N = 10) had small sample sizes, but I kept them in the analysis because the distributions of the scores in these group were symmetric with insignificant Shapiro-Wilk test.

With respect to the course delivery mode, the distributions of the scores were almost symmetric on all exams, except S1UPE in face-to-face group, the distribution of scores of which was slightly left-skewed with two mild outliers of 14.2% and 16%.

Shapiro-Wilk test was significant in all course delivery modes on all exams (*p*s < .001). The skewness and kurtosis ranged from -1.4 to -0.3. Thus, the normality assumption was not violated in all groups.

Levene's test of homogeneity was not significant on all exams (S1PE: *p* = .48; S1UPE: *p* = .32; S2PE; S1UPE: *p* = .17) except S2UPE (*p* =.02). But, on S2UPE, the ratio of the biggest SD of 22.23 to the smallest SD of 16.44 was about 1.4, indicating that the homogeneity assumption was not violated. Therefore, all assumptions needed for the full mixed ANOVA were met; I ran the test in SPSS.

Results of Full Mixed ANOVA

According to the ANOVA results, the order effect was significant with a small effect size ($F(1, 523) = 18.270$, $p < .001$, $\eta^2 = .034$), which paralleled the RQ2 analysis results. The order*mode interaction effect was not significant ($F(2, 522) = 1.548$, $p = .214$, $\eta^2 = .006$), suggesting that the order effect manifested similarly across all course delivery modes. The order*instructor interaction effect was significant with a small effect size ($F(6, 518) = 3.086$, $p = .006$, $\eta^2 = .035$), indicating that the order effect manifested differently across instructors. Similar to the results of RQ2 analysis, the order*format interaction had a medium significant effect ($F(1, 523) = 73.19$, $p < .001$, $\eta^2 = .124$). The interaction order*format*mode had a small significant effect ($F(2, 516) = 5.624$, $p = .004$, $\eta^2 = .021$), indicating that significant order*format interaction manifested differently across the course delivery modes. The order*format*instructor also had a small significant effect ($F(6, 518) = 5.608$, $p < .001$, $\eta^2 = .061$), suggesting that the significant order*format interaction manifested differently with respect to the instructors. Other results of this full mixed ANOVA paralleled the corresponding findings in RQ1 Set 2, RQ3 Set 2, RQ4 analyses: the format effect was significant ($F(1, 524) = 35.44$, $p < .001$, $\eta^2 =.064$), the format*mode effect was not significant ($F(2, 516) = 2.077$, $p = .126$, $\eta^2 = .008$), and the format*instructor effect was significant ($F(6, 516) =3.689$, $p = .001$, $\eta^2 =.041$). The summary of the full mixed ANOVA results is presented in Table 22.

Table 22

Results of Full Mixed ANOVA

Effect	*N*	*F*	*p*	η^2
order	525	18.27	.00	.03
order*mode		1.55	.21	.01
order*instructor		3.09	.01	.04
order*format		73.19	.00	.12
order*format*mode		5.62	.00	.02
order*format*instructor		5.61	.00	.06
format		35.44	.00	.06
format*mode		2.08	.13	.01
format*instructor		3.69	.00	.04

Thus, the manifestation of the insignificant order effect was not significantly different across the course delivery modes, but was significantly different with respect to the instructors and formats. The insignificant order*format interaction did not manifest significantly different across the course delivery modes and instructors. Next, I analyzed the exam scores of students who took the second version of unproctored exams (V2 group).

Testing RQs for V2 Groups

The were 50 students who took the second version of the unproctored exam in Set 1 and 76 who completed the second version of the unproctored exam in Set 2. In the combined set of the participants who took all four exams, 105 took the second version of at least one unproctored exam, but only 13 completed Version 2 on both unproctored exams. I tested the format effect in V2 group in Set 1 and Set 2 and the order effect in the combined V2 group students who took all four exams. The course delivery mode and instructor effects were not tested with V2 students because of the small sample size: the number of students with respect to instructors ranged from 1 to 9, and there were only 2 hybrid and 1online students. In V2 groups, there were no students with extended test time or students who took both tests in Set 2 in unproctored format.

The descriptive statistics of the scores in Set 1 V2 group, Set 2 V2 group, and combined S1&S2 V2 group of the students who took all four exams is provided in Table 23.

Table 23

Descriptive Statistics of V2 Groups

Group	*N*	M_{S1PE}*(SD)*	M_{S1UPE}*(SD)*	M_{S2UPE}*(SD)*	M_{S2PE}*(SD)*
S1V2	50	63.97 (19.95)	61.12 (21.07)	-	-
S2 V2	76	-	-	55.94 (19.54)	65.83 (17.72)
S1&S2V2	13	67.24 (21.34)	61.26 (20.65)	60.18(20.13)	67.21 (18.97)

Note: M_{S1PE} = the mean score on Set 1 proctored exam in %; M_{S1UPE} = the mean score on Set 1 unproctored exam in %; M_{S2UPE} = the mean score on Set 2 unproctored exam in %; M_{S2PE} = the mean score on Set 2 proctored exam in %; S1V2 = Set 1 version 2 group; S2V2 = Set 2 version 2 group; S1&S2V2 = combined Set 1 and Set 2 version 2 group.

According to Table 23, in Set 1 V2 group, the mean score on the proctored exam was 2.9% higher than on the unproctored exam. In Set 2 V2 group, the mean score on the proctored exam was 9.9% higher than on the unproctored exam. For comparison, in Set 1 main group, the mean score on the proctored exam of 67.26% was 1.3% lower than the mean score of 68.56% on the unproctored exam. In Set 2 main group, the mean score of 69.51% on the proctored exam was 5.7% higher than the mean score of 63.83% on the unproctored exam. In the combined Set 1 and Set 2 V2 group, the mean score on S1PE was 5.98% higher than on S1UPE, and the mean score on S2PE was 7% higher than on S2UPE. In the corresponding combined Set1 and Set 2 main group, the mean score of 71.73% on S1PE was just 0.1% lower than the mean score of 71.86% on S1UPE; the mean score of 70.22% on S2PE was 6.0% higher than on S2UPE. On all exams, the mean scores in the main groups were higher than the mean scores in the V2 groups.

According to the corresponding boxplots, in all V2 groups the distributions of the scores on all exams were almost symmetric without outliers. Shapiro-Wilk test was not significant in all V2 groups (S1PE: *p* =.27; S1UPE: *p* = .36; S2PE: *p* =.34; S2UPE: *p* = .22; Combined group: S1PE: *p* =.20; S1UPE: *p* = .43; S2PE: *p* =.30; S2UPE: *p* = .51). The normality assumption was not violated in all V2 groups, including the combined group with the small sample size of 13 students. I kept this group in the analysis. To test RQ1 hypothesis in V2 groups, I ran one-way repeated-measures ANOVA with V2 Set 1 scores and V 2 Set 2 scores separately. To test RQ2 hypothesis in V2 groups, I conducted two-way repeated measures ANOVA.

Results of Testing the Null Hypotheses of RQ1 and RQ2 in V2 groups

In V2 Set 1, the format effect was not significant (*F*(1,49) = .64, *p* =.43), while in V2 Set 2 a large significant format effect was observed (*F*(1,75) = 33.42, *p* <.001, η^2 = .31).Therefore, the null hypothesis for RQ1 was retained in V2 Set1, but rejected in V2 Set 2: the students’ scores on proctored and unproctored exams were not significantly different in Set 1, but significantly different in Set 2. The same results were obtained in the main group: the format effect was not significant in Set 1 group (*p* = .12), but significant in Set 2 group (*p* <.001, η^2 =. 15). However, the effect size was twice bigger in V2 group than in the main group. The results of RQ1 analysis in V2 groups are summarized in Table 24.

Table 24

Results of RQ1 Analysis in Set 1 and Set 2 V2 Groups

Group	*N*	*F*	*p*	η^2
Set 1 V2 Group	50	0.64	.430	.004
Set 2 V2 Group	76	33.42	.004	.310

In the combined Set 1 and Set 2 V2 group, a large significant order effect was observed (*F*(1,12) = 7.42, *p* =.02, η^2 = .38), but there were not significant format (*F*(1,12) = 0.07, *p* =.80) and order*format interaction effects (*F*(1,12) = 0.02, *p* =.89). The insignificant order*format interaction effect indicated that significant order effect did not manifest significantly differently across the formats. The null hypothesis of RQ2 in the V2 combined group was rejected. There was a significant difference in students' scores on proctored and unproctored exams with respect to the order in which the exams were administered. In the main group, the order effect was significant with a much smaller effect size (*p* <.001, η^2 = .08). Unlike V2 combined group, format and order*format interaction effects in the main group were significant (both *p* <.001) with the medium effect size of η^2 =.11 and η^2 =.13 respectively. The summary of the RQ2 results in the combined V2 group are provided in Table 25. Next, I analyzed the scores of students who had extended test time.

Table 25

Results of RQ2 Analysis in Combined V2 Groups

Effect	*N*	*F*	*p*	η^2
order	13	7.42	.02	.04
format		0.07	.80	.01
order*format		0.02	.89	.04

Testing RQs in Extended Test Time Groups

In Set 1, there were 16 students who took both exams with extended test times. In Set 2, 13 students had extended test times on both exams, and in the combined group, 10 students completed all four exams with extended time. In all these groups, there were no students who took the second version of the unproctored exams and students who took both exams in Set 2 in unproctored format. In spite of the small number of the students with the extended time accommodation, to examine whether the format and order effects in the extended time groups (Ext. time groups) and the main group had similar tendencies, I conducted RQ1 analysis in Set 1 Ext. time and Set 2 Ext. time and RQ2 analysis in the combined Ext. time group. The descriptive statistics of the scores in Set 1 Ext. time group, Set 2 Ext. time group, and combined S1&S2 Ext. time group of the students who took all four exams is provided in Table 26.

Table 26

Descriptive Statistics of Ext. Time Groups

Group	*N*	M_{S1PE}*(SD)*	M_{S1UPE}*(SD)*	M_{S2UPE}*(SD)*	M_{S2PE}*(SD)*
S1Ext. time	16	57.43 (17.70)	63.79 (20.55)	-	-
S2Ext. time	13	-	-	62.77 (17.10)	68.71 (18.52)
S&S2Ext.time	10	61.92 (18.86)	72.22 (15.62)	60.21(17.28)	67.34 (20.82)

Note: M_{S1PE} = the mean score on Set 1 proctored exam in %; M_{S1UPE} = the mean score on Set 1 unproctored exam in %; M_{S2UPE} = the mean score on Set 2 unproctored exam in %; M_{S2PE} = the mean score on Set 2 proctored exam in %; S1Ext.time = Set 1 extended time group; S2Ext.time = Set 2 extended time group; S1&S2Ext.time = combined Set 1 and Set 2 extended time group.

According to Table 26, in Set 1 Ext. time group, the mean score on the proctored exam was 6.4% lower than on the unproctored exam. In Set 2 Ext. time group, the mean score on the proctored exam was 5.9% higher than on the unproctored exam. For comparison, in Set 1 main group, the students had 1.3 % lower mean score on the proctored exam, and, in Set 2, the mean score on the proctored exam was 5.7% higher than on the unproctored. Although the means on Set 1 and Set 2 exams in the main group were higher than the corresponding means in the Ext. time groups (Main group: M_{S1PE}= 67.26%; M_{S1UPE} = 68.56%; M_{S2UPE} = 63.83%, M_{S2PE}= 69.51%), and the difference between the mean scores on Set 1 exams were higher in the Ext. time group, in Set 2, the differences between the mean scores in Set 2 in the main and Ext. time groups were very close. In the combined Set 1 and Set 2 Ext. time group, the mean score on S1PE was 10.3% lower than on S1UPE, and the mean score on S2PE was 7.1% higher than on S2UPE. Thus, in the combined Ext. time group, the students performed better on the exams that were administered second.

According to the corresponding boxplots, in Set 1 and Set 2 Ext. time group, the distribution of the scores was slightly left-skewed on S1PE and slightly right-skewed on S1UPE without outliers. Shapiro-Wilk test was not significant in both sets (S1PE: *p* =.87; S1UPE: *p* = .09; S2PE: *p* =.16; S2UPE: *p* = .75). In the combined Ext. time set, on S1PE, the distribution of scores was symmetric without outliers, on S1UPE, the distribution was right-skewed with two outliers of 40% and 46%. In the same group, the distribution was right-skewed on S2UPE and left-skewed on S2PE without outliers. Shapiro-Wilk test was not significant on all exams (S1PE: *p* =.86; S2PE: *p* =.89; S2UPE: *p* = .20), except S1UPE (*p* = .045). The skewness and kurtosis of the distribution of the scores on S1UPE were in acceptable limits (*skewness* = -1.3; *kurtosis* = -0.89). The normality assumption was not violated; I ran the tests in SPSS.

Results of Testing the Null Hypotheses of RQ1 and RQ2 in Ext. Time Groups

In Set 1 and Set 2 Ext. time group, the format effect was not significant (Set 1: *F*(1,15) = 1.84, *p* =.20; Set 2: *F*(1,12) = 2.18, *p* =.17).Therefore, the null hypothesis for RQ1 was retained in both sets: the students' scores on proctored and unproctored exams were not significantly different in Set 1 and Set 2 Ext. time groups. In comparison, in the main group, the format effect was not significant in Set 1 (*p* = .12), but a significant difference between the scores was observed in Set 2 (*p* <.001, η^2 =. 15). The findings of RQ1 analysis in Set 1 and Set 2 Ext. time groups are summarized in Table 27.

Table 27

Results of RQ1 Analysis in Set 1 and Set 2 Ext. Time Groups

Group	*N*	*F*	*p*	η^2
Set 1 Ext. Time Group	16	1.842	.198	.004
Set 2 Ext. Time Group	13	2.179	.173	.001

In the combined Ext. time group, the order, format, and order*format interaction effects were not significant (order: $F(1,9) = 0.18$, $p = .69$; format: $F(1,9) = 3.86$, $p = .08$; order*format: $F(1,9) = 0.83$, $p = .39$). Thus, the null hypothesis of RQ2 was retained: there was no difference between the scores of the students with extended test time on the proctored and unproctored exams with respect to the order in which exams were administering. Next, I analyzed the scores of the students who took both exams in Set 2 in unproctored format (UP group).

Testing Differences in Exam Scores in Unproctored Group

As I mentioned previously, during the data screening process, I found that 55 students took the second exam in Set 2, which had to be proctored, in unproctored format. Out of these 55 students, 51completed both exams in Set 2: the first one was synchronous unproctored exam and the second one was asynchronous unproctored exam. For this reason, I did not use the scores of these 51 students to test the study's hypotheses. Instead, I examined whether there was a significant difference in scores of the 51 students on these two exams.

According to the descriptive statistics analysis, in UP group, the students performed better on the second test in Set 2, which was administered asynchronously ($M_{Synch} = 59.07\%$, $SD_{Synch} = 18.62\%$, $M_{Asynch} = 68.87\%$, $SD_{Asynch} = 20.39\%$) with the difference in means of 9.80%. For comparison, the difference in scores on the same exams in the main group was 5.69%. The distributions of the scores on both exams were left-skewed, without outliers. Shapiro-Wiki test was significant for the scores of both exams (S2PE: p = .019, S2UPE: p = .045), suggesting deviations from normality. However, ANOVA is robust to deviations from normality in samples with sizes of at least 30 (Field, 2013). I ran a repeated measures ANOVA, according to which the difference between the scores was significant with a large effect size (F(1, 50) = 15.42, p =. 004, η^2 =.24). The corresponding 95% for the difference of the scores on the exams excluded 0% (CI [4.79, 14.83]), which supported the significant difference. Possible interpretations of this result are discussed in Chapter 5.

Testing the Hypotheses of RQs with Six Questions Scores

In most analyses described above, the mean scores on proctored and unproctored exams were not significantly different in Set 1, but, in Set 2, the mean score on the unproctored exam was significantly lower than on the proctored one. As I explained previously, Set 1 and Set 2 covered slightly different concepts. However, six out of 23 questions were identical across all four exams. To determine whether similar relationships between the study's variables were present in the subsets of the six identical questions, I calculated the average score of the six questions for each student who took all four exams. On these questions, a student could earn the maximum average score of 1.33 points. Then, I tested the study's hypotheses with the average scores of the six questions.

The descriptive statistics of the mean scores of the six questions in the main group of 525 students is provided in Table 28.

Table 28

Descriptive Statistics of the Scores of Six Identical Questions

Group	*N*	M_{S1PE}*(SD)*	M_{S1UPE}*(SD)*	M_{S2UPE}*(SD)*	M_{S2PE}*(SD)*
Main Group	525	.81 (.33)	.90 (.36)	.86 (.35)	.98 (.33)

Note: M_{S1PE} = the mean score of the six questions on Set 1 proctored exam; M_{S1UPE} = the mean score of the six questions on Set 1 unproctored exam; M_{S2UPE} = the mean score of the six questions on Set 2 unproctored exam; M_{S2PE} = the mean score of the six questions on Set 2 proctored exam.

According to Table 29, in both sets, the students perform better on the exam which was administered second. The scores increased from the first to the second exam by 0.09 points, decreased from the second to the third exam by 0.04 points, and increased again from the third to the fourth exam by 0.12 points. Although there was a decrease in scores between the second and third exam administration, the scores on the third exam were 0.05 points higher than on the first exam. Similar to the main group analysis with all questions, on the six identical questions, the students had higher scores on the unproctored exams when the unproctored exam was administered second and lower scores on the unproctored exam than on the proctored exam when the unproctored exam was administered first. The SDs were similar across all four exams. Next, I analyzed the scores of the six questions across the course delivery modes. The descriptive statistics of the scores on the six questions with respect to the course delivery mode is presented in Table 29.

Table 29

Descriptive Statistics of Scores on Six Questions Across the Course Delivery Modes

Mode	*N*	M_{S1PE}*(SD)*	M_{S1UPE}*(SD)*	M_{S2UPE} *(SD)*	M_{S2PE} *(SD)*
f2f	407	0.81 (0.33)	0.90 (0.35)	0.85 (0.35)	0.98 (0.32)
hyb	75	0.83 (0.29)	0.86 (0.35)	0.84 (0.35)	0.99 (0.35)
onl	43	0.81 (0.35)	0.94 (0.38)	0.90 (0.36)	0.99 (0.32)

Note: M_{S1PE} = the mean score of the six questions on S1PE; M_{S1UPE} = the mean score of the six questions on S1UPE, M_{S2PE} = the mean score of the six questions on S2PE; M_{S2UPE} = the mean score of the six questions on S2UPE.

As seen in Table 29, the scores of the students in all course delivery modes increased from the first to the second exam, decreased from the second to the third exam, and increased again from the third to the fourth exam. Thus, in both sets, the students' scores were higher on the exam that was administered second. Similar to the comparison of the scores across the course delivery modes in the main group with all exam items, on the six questions, the hybrid students had lower scores on both unproctored exams than students in face-to-face and online courses. The scores on the proctored exams were similar across the course delivery modes. With respect to the course delivery modes, the ratio of the biggest SD to the smallest SD on each exam was less than two (S1PE: 0.35/0.29 = 1.2; S1UPE: 0.38/0.35 = 1.1; S2UPE: 0.36/0.35 = 1.0; S2PE: 0.35/0.32 = 1.1). Next, I looked at the descriptive statistics of the scores of the six questions with respect to the instructors. These statistics are presented in Table 30.

Table 30

Descriptive Statistics of Scores on Six Questions Across the Instructors

Inst	*N*	M_{S1PE}*(SD)*	M_{S1UPE}*(SD)*	M_{S2UPE}*(SD)*	M_{S2PE}*(SD)*
I_1	12	0.84 (0.29)	1.00 (0.29)	0.96 (0.37)	0.91 (0.33)
I_2	68	0.85 (0.35)	0.89 (0.36)	0.81 (0.35)	0.99 (0.35)
I_3	270	0.79 (0.32)	0.89 (0.36)	0.89 (0.35)	0.97 (0.33)
I_4	43	0.76 (0.32)	0.93 (0.31)	0.79 (0.38)	0.97 (0.30)
I_5	72	0.77 (0.30)	0.88 (0.39)	0.83 (0.36)	1.00 (0.32)
I_6	50	0.96 (0.38)	0.99 (0.35)	0.84 (0.34)	1.05 (0.30)
I_7	10	0.95 (0.25)	1.05 (0.40)	0.76 (0.32)	1.05 (0.34)

Note: M_{S1PE} = mean score of the six questions on S1PE, M_{S1UPE} = mean score of the six questions on S1UPE; M_{S2PE} = mean score of the six questions on S2PE, M_{S2UPE} = mean score of the six questions on S2UPE ; Inst = instructor

According to Table 30, the students of all instructors had higher scores on the second exam in both sets except Instructor 1, whose students had the highest scores on the unproctored exam in Set 2, which was administered first. Thus, across all instructors, excluding Instructor 1, similar to the pattern observed in previous descriptive statistics analyses, the average scores on the six questions increased between the first and second exam, decreased between the second and third exam, and increased again between the third and fourth exam. With respect to the instructors, the ratio of the biggest SD to the smallest SD on each exam was less than two (S1PE: 0.38/0.25 = 1.5; S1UPE: 0.40/0.29 = 1.4; S2UPE: 0.37/0.32 = 1.2; S2PE: 0.35/0.30 = 1.2). Next, I evaluated the assumptions needed to test the study’s hypotheses.

According to the boxplots, the distribution of the scores of the six questions was symmetric on S1PE and S2PE and slightly left-skewed on S1UPE and S2UPE. There were no outliers in all distributions. Shapiro-Wilk test was significant in all distributions (*ps* < 0.001); skewness and kurtosis were in acceptable limits (S1PE: *skewness* = 0.16, *kurtosis* =- 0.86; S1UPE: *skewness* =- 0.21, *kurtosis* =- 1.09; S2UPE: *skewness* =- 0.82, *kurtosis* =- 0.39).

With respect to the course delivery modes, the distributions of the scores of the six questions were symmetric on S1PE and slightly left-skewed on all other exams. There were no outliers in all distributions. Shapiro-Wilk test was significant in all course delivery modes (S1PE: p_{f2f} <.001, p_{hyb} = .002, p_{onl} = .002; S1UPE: p_{f2f} <.001, p_{hyb} <.001, p_{onl} <.001; S2UPE: p_{f2f} <.001, p_{hyb} <.001, p_{onl} = .001; S2PE: p_{f2f} <.001, p_{hyb} <.001, p_{onl} <.001). Skewness and kurtosis were in acceptable limits (S1PE: *skewness f2f* = 0.15, *kurtosis f2f* = -0.88; *skewness hyb* = 0.18, *kurtosis hyb* = -0.55, *skewness onl* = 0.20, *kurtosis onl* = -0.99; S1UPE: *skewness f2f* = -0.24, *kurtosis f2f* = - 1.00; *skewness hyb* = 0.04, *kurtosis hyb* = - 1.29, *skewness onl*= -0.41, *kurtosis onl* = - 1.33; S2UPE: *skewness f2f* = -0.08, *kurtosis f2f* = - 1.02; *skewness hyb* = -0.17, *kurtosis hyb* = -1.09, *skewness onl*= -0.24, *kurtosis onl* = -0.91; S2PE: *skewness f2f* = -0.29, *kurtosis f2f* = -0.91; *skewness hyb* = -0.79, *kurtosis hyb* = - 0.32, *skewness onl*= -0.43, *kurtosis onl* = - 1.03).

With respect to the instructors, on S1PE, all distributions were left-skewed except Instructor 1and Instructor 4 group, which had right-skewed distributions. Instructor 5 group had symmetric distribution with four outliers. Instructor 2 group had right-skewed distribution on S2PE without outliers. All other distributions were left-skewed without outliers. Shapiro-Wilk test was not significant in Instructor 1 group (*p* =.128) and Instructor 7 (*p* =.854) group on S1PE, in Instructor 1 group on S1UPE (*p* =.169) and in Instructor 1 group on S2UE (*p* =.053), but significant in all other groups on all exams (p_s <.001). Skewness and kurtosis were in acceptable limits and ranged from -1.29 to -.03.

The Levene's test was not significant for the average scores of the six questions on S2PE (*p* =.81), S1UP (*p* =.48) and S2UPE (*p* =.76), but significant on S1PE (*p* =.02). However, the ration of the biggest SD to the smallest SD on S1PE was less than two (S1PE: 0.38/0.25 = 1.5). Thus, the homogeneity assumption was not violated. I conducted a repeated measures ANOVA with the scores of the six questions in Set 1 and Set 2 and then a mixed ANOVA with all study's variables.

Results of Testing the Null Hypotheses of all RQs with Six Questions Scores

On the six questions, there were a small significant format effect in Set 1 (F (1, 524) = 44.49, p <.001, η^2 = .08) and medium significant format effect in Set 2 (F (1, 524) = 66.63, p <.001, η^2 = .11). Thus, the null hypothesis of RQ1 was rejected. For comparison, in the same group of students, in the analysis with all questions, there was no significant format effect in Set 1 (F (1, 524) = .05, p = .82, η^2 <.001), but a medium significant format effect was observed in Set 2 (F (1, 524) = 104.71, p <.001, η^2 = .17). Then, I ran the mixed ANOVA with all involved variables.

According to the mixed ANOVA, the format effect was significant (F (1, 516) = 52.312, p <.001, η^2 =.09), but the order effect was not significant (F (1, 516) = 3.774, p = .053, η^2 =.007), which paralleled to the descriptive statistics observation that in both sets the average scores on the six questions were higher on the exams that were administered second. With the scores of the six questions, the null hypothesis of RQ2 was retained. For comparison, in mixed ANOVA analysis with all exam questions, both order and format effects were significant (format: F (1, 516) = 35.44, p <.004, η^2 =.06; F (1, 516) = 18.27, *p* <.001, η^2 =.03). The order*mode interaction effect with the scores on the six questions was significant with a small effect size *F* (2, 516) = 4.063, *p* = .018, η^2 =.016), indicating that the insignificant order effect manifested differently across the course delivery modes. The descriptive statistics corresponded to this fact: the hybrid students had smaller difference in scores between exams than face-to-face and online students when the unproctored exam was administered second and bigger difference in scores than face-to-face and online students when the unproctored exam was administered first (Set 1: diff f2f = .09, diff hyb = .03, diff onl = .13; Set 2: diff f2f = .13, diff hyb = .15, diff onl = .09). The order*format interaction effect with the scores of the six questions was not significant (*F* (1, 516) = 1.828, *p* = .177, η^2 =.004), while a small significant order*instructor interaction effect was observed (*F* (6, 516) = 3.394, *p* = .003, η^2 =.038), suggesting that the insignificant order effect manifested differently across the instructors, which paralleled to the descriptive statistics analysis. Both format*instructor and format*mode interaction effects with the six question were not significant (format*instructor: *F* (6, 516) = 1.399, *p* = .213, η^2 =.016); format*mode: (*F* (2, 516) = .232, *p* = .793, η^2 =.001), indicating that the significant format effect manifested similar across the instructors and course delivery modes. With the scores of the six questions, the null hypotheses of RQ3 and RQ4 were retained. For comparison, in the analyses with all questions, the format*mode effect was not significant (*F* (2, 516) = 2.08, *p* = .13, η^2 =.01), while the format*instructor effect was significant (*F* (6, 516) = 3.69, *p* <.001, η^2 =.04).

On the six questions, the order*format*mode interaction effect was not significant (*F* (2, 516) = 0.679, *p* = .508, η^2 =.003), while the order*format*instructor interaction effect was significant (*F* (6, 516) = 3.202, *p* = .004, η^2 =.036), suggesting that the insignificant order*format interaction effect manifested similarly across the course delivery modes but distinctly across the instructors. The findings of the mixed ANOVA with the six questions and the results of the mixed ANOVA with all questions are summarized in Table 31. Interpretations of these results are given in Chapter 5.

Table 31

Results of Full Mixed ANOVAs with Six Questions and All Questions

Effect	*N*	*F6*	*p6*	η^2 *6*	*Fall*	*pall*	η^2 *all*
order	525	3.774	.053	.007	18.270	.004	.030
order*mode		4.063	.018	.018	1.550	.210	.010
order*instructor		3.394	.003	.038	3.090	.010	.040
order*format		1.828	.177	.004	73.190	.004	.120
order*format*mode		0.679	.508	.003	5.620	.004	.020
order*format*instructor		3.202	.004	.036	5.610	.004	.060
format		52.312	.004	.092	35.440	.004	.060
format*mode		0.232	.793	.001	2.080	.130	.010
format*instructor		1.399	.213	.016	3.690	.004	.040

Note: F6 = F-statistics in the six questions analysis; *p6* = p-value in the six questions analysis; η^2 *6* = eta squared in the six questions analysis; *Fall* = F-statistics in all questions analysis; *pall* = p-value in all questions analysis; η^2 *all* = eta squared in all questions analysis.

Testing Differences in GPA and Age across the Modes and Instructors

In the previous analyses, some significant differences across the course delivery modes and instructors' groups were identified. To understand whether these distinctions could be attributed to differences in academic abilities and age, I conducted two one-way ANOVAs with the variables GPA and age. The total set of all 850 students was used for this analysis.

The descriptive statistics of the participants' GPA and age with respect to the course delivery mode is presented in Table 32.

Table 32

Descriptive Statistics of GPA and Age across Course Delivery Modes

Mode	*N*	*M_{GPA} (SD)*	*M_{Age} (SD)*
f2f	702	3.15 (0.53)	21.59 (4.35)
hyb	92	3.13 (0.58)	25.17 (6.75)
onl	55	3.23 (0.55)	21.98 (4.57)

Note: M_{GPA} = GPA mean; M_{Age} = age mean.

According to Table 32, the GPA means and SDs were similar in all course delivery modes. The mean and SD of the variable age were close in the face-to-face and online groups, but higher in the hybrid group. The descriptive statistics of GPA and age with respect to instructors is shown in Table 33.

Table 33

Descriptive Statistics of GPA and Age across Instructors

Instructor	*N*	*M*$_{GPA}$ *(SD)*	*M*$_{age}$ *(SD)*
I_1	20	3.08 (0.57)	21.65 (7.05)
I_2	109	3.15 (0.51)	21.30 (3.76)
I_3	330	3.17 (0.54)	22.58 (5.42)
I_4	80	3.25 (0.52)	20.98 (3.21)
I_5	153	3.06 (0.54)	21.95 (4.66)
I_6	80	3.14 (0.56)	21.53 (3.13)
I_7	78	3.18 (0.53)	22.47 (5.69)

Note: M$_{GPA}$ = GPA mean; *M*$_{Age}$ = age mean; I = instructor.

According to Table 33, the GPA means and SDs did not differ a lot across the instructors, except Instructor 5 group, in which the GPA standard deviation was about two points higher than in the groups of other instructors. The average age was about 1.5 years higher in Instructor 3 group than in the groups of all other instructors. The SDs ranged from 3.1 years to 7.1 years. Next, I evaluated the statistical assumptions.

The boxplots of the GPA distributions with respect to the course delivery modes were symmetric. One outlier of 1 was observed in the face-to-face group, and one outlier of 1.3 was present in the hybrid group. The distribution of the GPAs in the online group did not have outliers. Shapiro-Wilk's test was not significant in the online group (*p* = .051), but significant in the face-to-face (*p* <.001) and hybrid groups (*p* = .006), indicating deviation from normality. However, ANOVA is robust to nonnormality if the sample size is bigger than 30 (Field, 2013), as it was in all course delivery mode groups. The Levene's test for GPA across the course delivery modes was not significant, indicating no significant difference in SDs across the course delivery modes (*p* = .344).

The boxplots of the GPA distributions with respect to the instructors were also approximately symmetric. Instructor 2 group had three outliers of 1.90, 1.91, and 2, Instructor 3 group had one outlier of 1.29, and Instructor 5 group had an outlier of 1.0. There were no outliers in all other instructors' groups. Shapiro-Wilk's test was not significant in Instructor 1 group (*p* = .507), but significant in all others (I2: *p* = .038; I3: *p* <.001; I4: *p* = .006; I5: *p* = .004; I6: *p* = .008; *p* = .021). The sample size in all instructors' groups was bigger than 30 except Instructor 1 group (*N* = 20), which had insignificant Shapiro-Wilk test. The Levene's test for GPA across the instructors was also not significant (*p* = .915), suggesting that variances were not significantly different across the instructors.

The boxplot for the distribution of ages in the face-to-face group was symmetric, but had 12 outliers ranged from 32 to 50 years old. The boxplot for the hybrid group was slightly right-skewed, with three outliers of 43, 45, and 47 years old. The online group had almost symmetric distribution with three outliers of 15, 28 and 29 years old. Shapiro-Wilk's test was significant in all course delivery mode groups (*ps* <.001). The Levene's test also was significant (*p* <.001), but the ratio of the biggest SD to the smallest SD was less than two (6.80/4.35 = 1.56).

The distributions of ages across the instructors were right-skewed; the outliers were present in all groups. Shapiro-Wilk's test was significant across all instructors (*ps* < .001). The Levene's test was also significant (*p* <.001), but an alternative homogeneity Brown-Forsythe's test was not significant (*p* = .072). I ran one-way ANOVAs for GPA and age across the course delivery modes and instructors in SPSS.

Results for Testing GPA and Age Differences across Modes and Instructors

According to the ANOVA results, there were no significant differences in the participants' GPA across course delivery modes (*F*(2,847) = 0.748, *p* = .474, η^2 = .002) and instructors (*F*(6,843) = 1.313, *p* = .248, η^2 = .009). However, a significant difference in the students' age across the course delivery modes was determined (*F*(2,847) = 25.520, *p* <.001, η^2 = .057). To identify in which course delivery mode the difference occurred, I applied Bonferroni post hoc test and found a significant difference in age between hybrid and face-to-face students (*p* <.001) and between hybrid and online students (*p* <.001). Thus, the hybrid students were significantly older than face-to-face and online students (f2f: M_{Age} = 21.59; hyb: M_{Age} = 25.29; onl: M_{Age} = 21.98). There was no significant difference in the participants' age across the instructors (*F*(6,843) = 2.058, *p* = .056, η^2 = .014). Interpretations of these results are given in Chapter 5.

Testing Attrition due to Dropout

As I described previously, although there were students who did not take some or all exams in Set 1, but took some or all exams in Set 2, the number of examinees was decreasing with each exam administration. To understand whether the implementation of the proctored and unproctored web-based exams by the department influenced attrition due to drop out, I compared the attrition rate before and after the implementation was incorporated. The descriptive statistics for the dropout rate during each period is provided in Table 34.

Table 34

Dropout Rate before and after Web-based Exams' Implementation

Period	Students Enrolled	Students Dropped	*N*	*M*	*SD*
Before	720	295	20	40.97	9.87
After	1154	362	33	31.37	13.44

Note: N = the number of stats sections, *M* = the dropout rate mean in %

According to Table 34, on average, the dropout rate decreased by almost 10% after the implementation of the web-based exams took place. The SDs decreased by 3.6%. Then, I examined the dropout rate with respect to the course delivery modes; this statistics is shown in Table 35.

Table 35

Dropout Rate before and after Implementation across the Modes

Period	Mode	*N*	*M*	*SD*
Before	f2f	17	41.08	10.25
	hyb	2	41.05	12.65
	onl	1	38.00	-
After	f2f	27	30.90	13.44
	hyb	4	35.70	14.37
	onl	2	34.07	14.04

Note: N = number of stats sections, *M* = dropout rate mean in %

As seen in Table 35, after the proctored and unproctored web-based exams were implemented, the dropout rate means decreased by 10.2% in the face-to-face sections, 5.4% in the hybrid sections, and 3.9% in the online sections. The increase in SDs ranged from 1.7% to 3.2%. To test whether the observed differences in the dropout rate were significant, I verified the needed assumptions and then conducted a one-way ANOVA followed by the mixed ANOVA.

According to the boxplots, the distributions of the dropout rates were almost symmetric before and after the exams were implemented. There was one outlier with the dropout rate of 65.8% after the implementation. Shapiro-Wilk's test was not significant before and after the incorporation of the exam (before: $p = .90$; after: $p = .16$). The Levene's test was also not significant ($p = .21$).

The distribution of the dropout rates was slightly right-skewed in the face-to-face group and slightly left-skewed in the hybrid and online groups. There were no outliers across all delivery modes. Shapiro-Wilk's test was not significant in all three course delivery modes (f2f: $p = .52$; hyb: $p = .63$; onl: $p = .57$). The Levene's test was also not significant ($p = .80$). Thus, all needed assumptions were met; I ran the tests in SPSS.

Results for Testing Attrition due to Dropout

There was a medium size significant difference in the dropout rate before and after the web-based exams were implemented ($F(1, 51) = 7.19$, $p < .001$, $\eta^2 = .12$). Thus, the decrease of the dropout rate after the implementation was significant. With respect to the course delivery modes, no significant difference in the dropout rate before and after the web-based exam incorporation was determined ($F(2, 47) = 0.09$, $p = .92$, $\eta^2 = .004$). Interpretations of this result are provided in Chapter 5. The period*mode interaction effect, was also not significant ($F(2, 47) = 0.15$, $p = .86$, $\eta^2 = .006$), indicating the changes in the dropout rate manifested similarly across the course delivery modes. The variable period stands for the period before and after the implementation.

Testing Reliability of Study's Exams

As I described in Chapter 3, the department established the content validity of the web-based exams, but no formal statistical analyses were done to test reliability and construct validity of the used tests. To verify whether the four exams were reliable, I conducted a reliability analysis in SPSS. The scores of all 850 students who took at least one study's exams were used for this analysis.

The following basic assumptions underline internal consistency reliability analysis: the responses of one subject should be independent from responses of any other subject for each administration of a test and number of subjects should be at least 10 times more than the number of items in the scale. It is recommended to exclude from reliability analysis all exams with missing item responses (Field, 2013; Tavakol, Dennik, & Tavakol, 2011). Cronbach alpha of at least .80 indicates reliable instrument for high-stake assessments and at least .70 for tests created by teachers (Ertürk, 2015; Field, 2013; Tavakol et al., 2011). Items of a reliable scale correlate with the total and have coefficients at least .3 (Field, 2013).

The responses of all students were independent of each other on all four exams. Each exam had 23 items; thus, at least 230 subjects were needed for adequate reliability analyses. On S1PE, there were 536 students who did not miss any items; on S1UPE, 439 participants responded to each test item. On V2 S1UPE, there were only 29 students out of 55 who answered all questions. On S2UPE, 444 students responded to all test items; while on the second version of this test 55 students out of 101 answered all questions. On S2PE, 635 students responded to all test items.

Results for Testing Reliability of Study's Exams

On S1PE, Cronbach alpha was .84, and the standardized alpha was .86. The corrected-item total correlation coefficients ranged from .3 to .6. Exclusion none of the items would increase Cronbach alpha of .84. Thus, S1PE had adequate reliability.

On S1UPE, the alpha coefficient was .76 with the standardized alpha of .79. The corrected-item total correlation coefficients ranged from .3 to .5, except items 2, 3, and 18, the coefficients of which were around .2. Deleting none of the items would increase the alpha coefficient of .76. S1UPE also had sufficient reliability. The second version of S1UPE had Cronbach alpha of .79 with the standardized alpha of .81.

On S2UPE, Cronbach alpha was .82 with the standardized alpha of .86. The corrected-item total correlation coefficients ranged from .3 to .6. Removing none of the items would increase Cronbach alpha of .82. S2UPE was reliable. The second version of S2UPE had the alpha coefficient of .83 with the standardized alpha of .86.

On S2PE, Cronbach alpha was .76 with the standardized alpha of .79. The corrected-item total correlation coefficients ranged from .3 to .5, except items 4, 5, and 20, which had coefficients around .2. Exclusion none of the items would increase Cronbach alpha of .76. S2UPE had adequate reliability. The reliability analysis findings are summarized in Table 36.

Table 36

Results of Reliability Analysis

Exam	*n*	Cronbach alpha	Standardized Cronbach alpha
S1PE	536	.84	.86
S1UPE	439	.76	.79
S2UPE	444	.82	.86
S2PE	635	.76	.79
S1UPE V2	29	.79	.81
S2UPE V2	55	.83	.86

Note: n = the number of students who responded to all test items.

Testing Construct Validity of Study's Exams

To evaluate the construct validity of exam scores, exploratory and confirmatory factor analyses can be used (Dundar, Temel, & Gunduz, 2016; Stapleton, 1997; Tavakol et al., 2011; Williams, Fall, Eaves, & Darch, 2007). Exploratory factor analysis is used to examine intercorrelations between exam items and combine them into groups (factors) that measure similar constructs. When the factor structure is identified, confirmatory factor analysis (CFA) can be utilized to measure the model fit. An adequate model fit suggests validity of the exam structure (Dundar et al., 2016; Williams et al., 2007).

For a reliable factor analysis several assumptions are required. The sample size should be at least 300, and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy should be at least .5. Additionally, the correlation between the factors should not be too high and too low. For this reason, Bartlett's test of sphericity should be significant, indicating that the correlation between the items is significantly different from zero, and the determinant of the correlation matrix should be greater than .00001, justifying low multicollinearity (Field, 2013). According to Dundar et al. (2016), Field (2013), and Tabachnick and Fidell (2007), the eigenvalues of at least one, loading coefficients of at least .3, and at least three items in each factor are recommended for extracting adequate factors.

Results for Testing Construct Validity of S1PE Scores

On S1PE, there were 536 exams with no missing responses. For the scores of these 536 exams, many coefficients in the correlation matrix were bigger than .3, and none of them exceeded .9. The determinant of the matrix was .01, much bigger than required .00001. The KMO index was .91 > .5; Bartlett's test of sphericity was significant (χ^2 (253) = 2450.70; $p < .001$). Therefore, factor analysis was appropriate for S1PE scores and had to yield distinct and reliable factors. I ran the Principal Component Analysis (PCA) in SPSS.

According to PCA, five eigenvalues of 5.73, 1.46, 1.14, 1.08, and 1.04 were bigger than one, identifying a five-factor model which accounted for 45.42% of the total variance. Out of this 45.42%, the first factor accounted for 24.89%, the second one for 6.36%, the last three for 4.96%, 4.70%, and 4.51% respectively. Although there were five eigenvalues greater than one, the scree plot had one inflection point at the second eigenvalue, suggesting two unique components (factors). When I tried to run the test with four or three factors, there were less than three entries with coefficients of at least .3 in some columns. Tavakol et al. (2011) recommended using eigenvalues of 1.25 to identify the number of factors. Taking into consideration all these facts, I decided to utilize a two-factor model with the total variance of 31.25%.

To interpret a factor structure, rotation is recommended (Field, 20013; Tabachnick & Fidell, 2007). There are two major types of rotation: orthogonal for not correlated factors and not orthogonal (oblique) for correlated ones (Field, 2013). Because all items on each exam were related to the same course, I assumed that the underlying factors could correlate to a certain degree and used promax rotation, an oblique rotation recommended by Field (2013) for big sample sizes.

After I ran the PCA with the two factors selected for an extraction, the pattern matrix had 16 items for the first factor and seven items for the second factor. As I described previously, on each exam, there were six cumulative questions. Thus, S1PE had 17 items on inferences with proportions and six items on sampling and experiment covered in previous units. After a short analysis of the two factors, I realized that 15 out of 16 items in the first factor were on inferences with proportions and five out of seven items in the second factor were on cumulative concepts about experiments and sampling. For this reason, I called first factor Proportions and the second one Experiment. Next, I tested reliability of each subscale. The alpha coefficient for the first factor was .84, which indicated high reliability. The alpha for the second factor was .59, which, according to Field (2013), could be considered as acceptable. The results of the PCA for the two-factor model are provided in Table 37.

Table 37

Results of PCA for Two-factor Model of S1PE Scores (n=536)

No.	Item	h^2	Proportion	Experiment
Q6	Calculate SD of the sampling distribution	.46	**.72**	-.09
Q15	Calculate 95% CI for pop. proportions	.43	**.70**	-.09
Q12	Calculate bin. distr. probability P(x >2)	.43	**.68**	-.05
Q11	Calculate bin. distr. probability P(x=2)	.46	**.67**	.01
Q21	Calculate test statistics	.35	**.63**	-.09
Q9	Calculate reasonably likely interval	.46	**.62**	.09
Q16	Calculate margin of error of the 95% CI	.34	**.56**	.04
Q17	Interpret margin of error	.23	**.55**	-.19
Q23	Identify graph of normal distribution	.23	**.49**	.12
Q13	Calculate normal distribution probability	.30	**.47**	-.03
Q20	Identify Ho and Ha for pop. proportion	.21	**.40**	.11
Q19	Identify symbol for pop. proportion	.26	**.40**	.18
Q22	Interpret hypothesis test results	.20	**.37**	.01
Q7	Calculate sample proportion	.30	**.37**	.25
Q14	Identify symbol of sample proportion	.20	**.36**	.11
Q8	Identify assumptions of normality	.20	**.35**	.12
Q4	Identify factors in an experiment	.45	-.11	**.72**
Q5	Identify response variable in experiment	.45	-.08	**.71**
Q2	Identify the most representative sample	.30	-.08	**.54**
Q3	Identify experimental units	.30	.08	**.50**
Q18	Interpret 95% CI for pop. proportion	.22	.01	**.46**
Q1	Identify type of sampling	.27	.10	**.46**
Q10	Identify binomial distribution	.30	.27	**.36**
Eigenvalues			5.73	1.08
Percent of variance			24.89	6.36
Alpha			.84	.59

Note: h^2 = communalities; No. = item number. Factor loading used to determine the factors appears in bold.

To test whether the two-factor model was a good fit for the data, I conducted Confirmatory Factor Analysis (CFA) by using Structural Equation Modeling technique in SPSS AMOS 24. It is recommended to test the fit of a model with multiple criteria: the likelihood ratio chi-square statistics χ^2 and its ratio with the degrees of freedom χ^2/df, the root mean square error approximation RMSEA, goodness-of-fit index GFI, adjusted goodness of fit index AGFI, comparative fit index CFI, and the standardized root mean square RMR (Dundar et al., 2016; Tavakol et al, 2011; Williams et al., 2007). When a model fit is acceptable, the chi-square test is insignificant or the ratio of χ^2/df is less than two, RMSEA is below .05, RMR is close to 0, GFI, AGFI, and CFI are greater than .9 (Dundar et al., 2016; Tavakol et al, 2011). According to Schumaker and Lomax (2004), the ratio of χ^2/df less than five is adequate in big samples.

I constructed a path diagram for the two-factor model and ran a maximum likelihood test in AMOS. The diagram with the corresponding coefficients is presented in Figure 4.

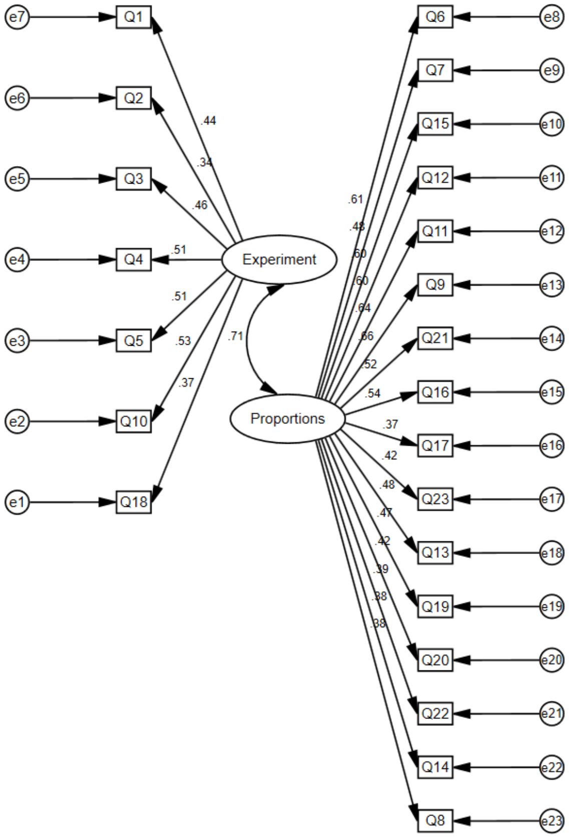


Figure 4. Two-factor model for S1PE.

According to Figure 4, the coefficients in the first factor ranged from .34 to .53; the coefficient in the second factor ranged from .37 to 0.64. The correlation between the two factors was .71. The results of the maximum likelihood test are provided in Table 38.

Table 38

Results of CFA for Two-factor Model of S1PE Scores (n =536)

Model	χ^2	*df*	*p*	χ^2/df	GFI	AGFI	CFI	RMR	RMSEA
2-factor	401.49	229	.004	1.75	.94	.93	.92	.03	.03

As seen in Table 38, the Chi-square test was significant (p < .05). However, the ratio χ^2/df was less than two; all indices were in acceptable limits, indicating a good fit of the model to the S1PE scores. Therefore, the construct validity of S1PE was adequate.

Results for Testing Construct Validity of S1UPE Scores

S1UPE, a parallel version of S1PE, was administered in an unproctored environment. I conducted the PCA and CFA with S1UPE data, including only the scores of unproctored exams that were taken synchronously and did not have missing items (*N* = 439). For this set of scores, many coefficients in the correlation matrix were bigger than .3 and less than .9. The determinant of the matrix was .046 > .00001. The KMO index was .84 > .5; Bartlett's test of sphericity was significant (χ^2 (253) = 1323.12; *p* < .001). Thus, factor analysis was appropriate for S1UPE scores and had to yield distinct and reliable factors. I ran PCA with promax rotation using S1UPE scores.

There were seven eigenvalues of 4.31, 1.42, 1.30, 1.18, 1.13, 1.05 and 1.02 bigger than one, suggesting a seven-factor model which accounted for 49.58% of the total variance. Out of this 49.58%, the first factor accounted for 18.74%, the second one for

6.12%, the last five for 5.65%, 5.13%, 4.93%, 4.55%, and 4.42% respectively. The scree plot had an inflection point at the second eigenvalue. I utilized the two-factor model, ran the test, and compare the models across Set 1 exams.

The factor loading was similar to the loading of S1PE analysis: 17 items loaded on the factor that was related to the inferences with proportions, and six items loaded on the factor that corresponded to the concepts of experiments. However, unlike PCA in S1PE, in the pattern matrix of which all items had coefficients bigger than .3, some items in the pattern matrix of S1UPE scores had coefficients between .2 and .3. I kept all items in the analysis of S1UPE scores and tested the reliability of each subscale in the two-factor model for this exam. The alpha coefficient for the first factor was .72, which indicated high reliability for tests created by instructors. The alpha for the second factor was .56, which suggested acceptable reliability. The results of the PCA for the two-factor model are provided in Table 39.

Table 39

Results of PCA for Two-factor Model of S1UPE Scores (n=439)

No.	Item	h^2	Proportion	Experiment
Q9	Calculate reasonably likely interval	.44	**.68**	-.04
Q15	Calculate 95% CI for pop. proportions	.35	**.61**	-.05
Q21	Calculate test statistics	.33	**.59**	-.05
Q11	Calculate bin. distr. probability P(x=2)	.24	**.52**	-.11
Q7	Calculate sample proportion	.25	**.52**	-.06
Q13	Calculate normal distribution probability	.28	**.50**	.05
Q16	Calculate margin of error of the 95% CI	.24	**.49**	.01
Q6	Calculate SD of the sampling distribution	.47	**.56**	.03
Q12	Calculate bin. distr. probability P(x >2)	.23	**.43**	.28
Q18	Interpret 95% CI for pop. proportion	.14	**.40**	-.16
Q22	Interpret hypothesis test results	.18	**.33**	.17
Q8	Identify assumptions of normality	.15	**.31**	.14
Q14	Identify symbol of sample proportion	.12	**.28**	.23
Q1	Identify type of sampling	.20	**.28**	.14
Q23	Identify graph of normal distribution	.12	**.25**	.16
Q17	Interpret margin of error	.07	**.22**	.10
Q2	Identify the most representative sample	.08	**.21**	.12
Q4	Identify factors in an experiment	.50	-.11	**.72**
Q3	Identify experimental units	.40	-.16	**.68**
Q5	Identify response variable in experiment	.41	-.06	**.66**
Q10	Identify binomial distribution	.30	.25	**.40**
Q19	Identify symbol for pop. proportion	.20	.26	**.27**
Q20	Identify Ho and Ha for pop. proportion	.17	.23	**.26**
Eigenvalues			4.31	1.42
Percent of variance			18.74	6.17
Alpha			.72	.56

Note: h^2 = communalities; No. = item number. Factor loading used to determine the factors appears in bold.

Next, I conducted CFA of the two-factor model with S1UPE scores in AMOS. The path diagram for this model is shown in Figure 5.

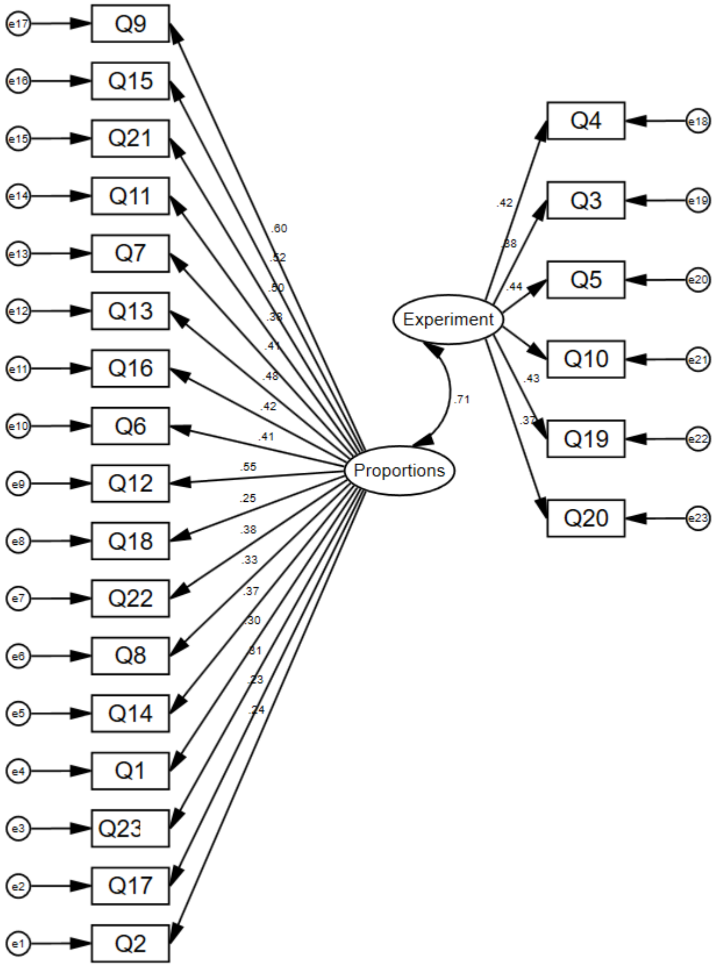


Figure 5. Two-factor model for S1UPE.

As seen in Figure 5, three coefficients in Proportions, .25 for item Q18, .23 for item Q17, and .24 for item Q1, were less than .3; other coefficients in this factor ranged from .30 to .60. All coefficients in Experiment were at least .3 and ranged from .37 to .44. In spite of a few differences in loading coefficients, the path diagrams for two-factor models for proctored and unproctored exams in Set 1 were similar. The correlation between the Proportions and Experiment factors was exactly the same, .71. The results of the maximum likelihood test for S1UPE are provided in Table 40.

Table 40

Results of CFA for Two-factor Model of S1UPE Scores (n =439)

Model	χ^2	*df*	*p*	χ^2/df	*GFI*	*AGFI*	*CFI*	*RMR*	*RMSEA*
2-factor	339.40	229	.004	1.48	.94	.92	.90	.03	.03

According to Table 40, on S1UPE, although the goodness-of-fit test was significant ($p <.05$), the chi-square and the ratio χ^2/df were slightly smaller than for S1PE (S1PE: χ^2 =401.49, χ^2/df =1.75; S1PE: χ^2 =339.40; χ^2/df =1.48). All other measures of the model fit on S1PE and S1UPE were almost identical (S1PE: *GFI* = .94; *AGFI* = .93, CFI = .92; *RMR* =.03; *RMSEA* =.03; S1UPE: *GFI* = .94; *AGFI* = .92, *CFI* = .90; *RMR* =.03; *RMSEA* =.03). Thus, on S1UPE, all indices were in acceptable limits, suggesting a good fit of the model to the S1UPE scores. The construct validity of S1UP exam was adequate. Additionally, the two-factor models suggested by PCA manifested very similarly on both exams of Set 1. Next, I conducted the PCA and CFA with S2UPE scores.

Results for Testing Construct Validity of S2UPE Scores

For the PCA and CFA for S2UPE, I also used the scores of the students who took the unproctored exam synchronously and did not have missing items (*N* = 444). For this set of scores, many coefficients in the correlation matrix were between .3 and .9. The determinant of the correlation matrix was .008 > .00001. The KMO index was .90 > .5; Bartlett's test of sphericity was significant (χ^2 (253) = 2096.57, $p < .001$). Therefore, factor analysis was appropriate for S2UPE scores and had to generate distinct and reliable factors. I ran PCA for S2UPE scores.

According to PCA for S2UPE, six eigenvalues of 5.78, 1.39, 1.20, 1.15, 1.05 and 1.03 were bigger than one, identifying a six-factor model which accounted for 50.47% of the total variance. Out of this 50.47%, the first factor accounted for 25.14%, the second one for 6.04%, the last four for 5.20%, 5.01%, 4.59% and 4.49% respectively. When I ran the test with the different number of factors, the six-factor and five-factor models had less than three entries with coefficients of at least .3 in some columns. The three-factor model had cross-loading items. The two-factor and four-factor models seemed suitable. However, the scree plot had an inflection point at the second eigenvalue, suggesting a two-factor model. I decided to utilize a two-factor model with the total variance of 31.19%.

For the two-factor model of S2UPE, the PCA with promax rotation generated 11 items in the first factor and 12 items in the second factor. Set 2 exams included topics on inferential procedures with means and the six identical questions. The 12 items in the first factor were on calculations and concepts with means. I called this factor Means. The second factor included five out of six identical questions and some items related to conceptual procedures with means. To be consistent with the factor analysis of exams in Set 1, I called the second factor Experiment. The alpha coefficient for the first factor was .72 and for the second factor .71, which indicated good reliability of both subscales. The results of the PCA for the two-factor model are provided in Table 41.

Table 41

Results of PCA for Two-factor Model of S2UPE scores (n=444)

No.	Item	h^2	Means	Experiment
Q14	Identify components of ANOVA table	.52	**.81**	-.19
Q21	Calculate F-statistics	.48	**.77**	-.13
Q11	Calculate t-statistics	.32	**.65**	-.17
Q19	Identify Ho and Ha for ANOVA	.46	**.60**	.10
Q15	Calculate 95% CI for means	.30	**.55**	.10
Q6	Identify ANOVA assumptions	.31	**.48**	.12
Q7	Calculate degree of freedom	.19	**.46**	-.05
Q13	Calculate normal distribution probability	.26	**.43**	.12
Q1	Identify statistical technique	.29	**.42**	.17
Q8	Identify CI assumptions	.22	**.37**	.14
Q23	Identify graph of F distribution	.21	**.28**	.23
Q2	Identify the most representative sample	.37	-.18	**.69**
Q22	Interpret HT vs CI results	.30	-.06	**.58**
Q4	Identify factors in an experiment	.35	.05	**.57**
Q12	Identify appropriate hypothesis test	.22	-.18	**.55**
Q18	Interpret 95% CI for means	.30	-.01	**.55**
Q5	Identify response variable in an experiment	.39	.15	**.53**
Q3	Identify experimental units	.29	.03	**.52**
Q17	Identify significance of a test	.37	.17	**.49**
Q9	Identify Ho and Ha for paired test	.45	.33	**.42**
Q16	Calculate the margin of error	.26	.14	**.42**
Q20	Identify Ho and Ha for paired test	.14	-.06	**.41**
Q10	Identify Type I and Type 2 errors	.22	.23	**.30**
Eigenvalues			5.78	1.39
Percent of variance			25.16	6.04
Alpha			.72	.71

Note: h^2 = communalities; No. = item number. Factor loading used to identify the components appears in bold.

Next, I conducted CFA of the two-factor model with S2UPE scores in AMOS. The path diagram for this model is shown in Figure 6.

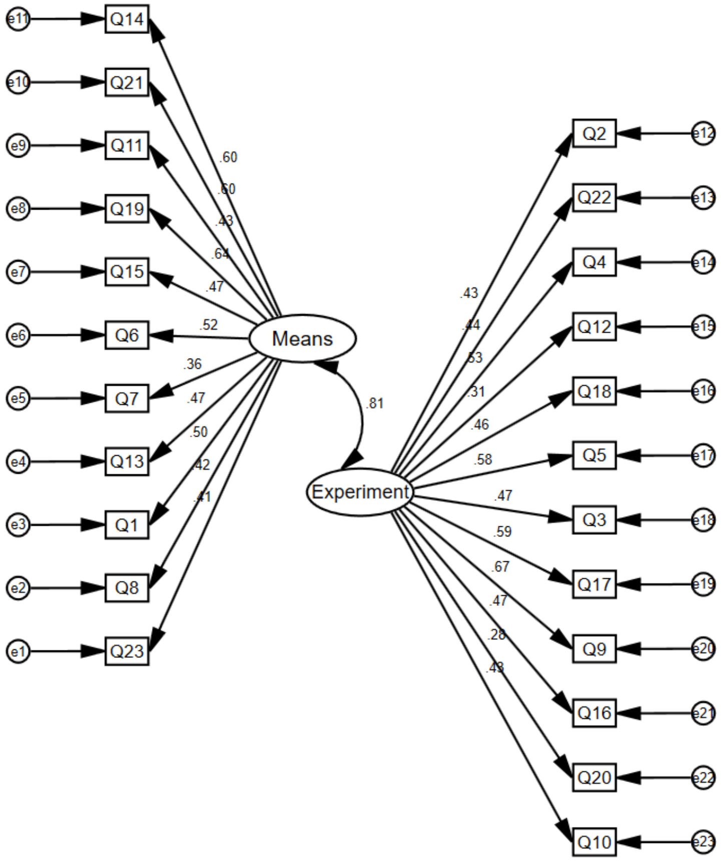


Figure 6. Two-factor model for S2UPE.

According to Figure 6, the coefficients in Means ranged from .36 to .64, and the coefficients in Experiment ranged from .28 to .67. The correlation between the Means and Experiment factors was .81. The results of the maximum likelihood test for S2UPE are provided in Table 42.

Table 42

Results of CFA for 2-Factor Model of S2UPE Scores (n =444)

Model	χ^2	df	p	χ^2/df	GFI	AGFI	CFI	RMR	RMSEA
2-factor	402.88	229	.004	1.76	.93	.91	.91	.04	.04

As seen in Table 42, on S2UPE, the chi-square test was significant (p <.05), but the ratio χ^2/df was 1.76 < 2. All indices were in acceptable limits, suggesting a good fit of the two-factor model to the S2UPE scores. Thus, the construct validity of S2UP exam was adequate. Next, I conducted the PCA and CFA with S2PE scores.

Results for Testing Construct Validity of S2PE Scores

The second exam in Set 2, S2PE, was a parallel version of S2UPE administered in proctored format. The scores of 635 students who did not miss any items were used for the PCA and CFA of S2PE. For this set of scores, many coefficients in the correlation matrix were bigger than .3 and less than .9. The determinant of the matrix was .04 > .00001; the KMO index was .83 > .5; Bartlett's test of sphericity was significant (χ^2 (253) = 2006.36, p < .001). Thus, factor analysis was appropriate for S2PE scores and had to generate distinct and reliable factors. I ran PCA for S2PE scores.

Seven eigenvalues of 4.28, 1.58, 1.41, 1.19, 1.10, 1.03 and 1.01 were bigger than one, identifying a seven-factor model which accounted for 50.36% of the total variance. Out of this 50.36%, the first factor accounted for 18.60%, the second one for 6.85%, the last five for 6.12%, 5.16%, 4.78%, 4.49%, and 4.36% respectively. On the corresponding scree plot, there was one inflection point at the second eigenvalue with the big change in concavity of the plot, suggesting suitability of the two-factor model. The seven-factor, six-factor, and five-factor models had less than three loadings in some columns; the four-factor model had cross loading items; the three-factor model had five items with loadings substantially less than .3, while the two-factor model had only two items slightly less than

.3. Based on all these facts, for S2PE, I decided to test the two-factor model with the total variance of 25.45%.

For S2PE scores, the PCA with promax rotation for the two-factor model generated nine items in the first factor and 14 items in the second factor. Similar to the S2UPE PCA, the nine items in the first factors were related to inferential procedures with means, and the questions on the experiment concepts were in the second factor. I called the first factor Means and the second one Experiment. The alpha coefficient for the first factor was .72 and for the second one .69, indicating acceptable reliability. The results of the PCA for the two-factor model are provided in Table 43.

Table 43

Results of PCA for Two-factor Model of S2PE Scores (n=635)

No.	Item	h^2	Means	Experiment
Q14	Identify components of ANOVA table	.40	**.70**	-.20
Q11	Calculate t-statistics	.29	**.59**	-.20
Q16	Calculate the margin of error	.30	**.58**	-.10
Q21	Calculate F-statistics	.32	**.57**	-.13
Q6	Identify ANOVA assumptions	.37	**.56**	-.17
Q15	Calculate 95% CI for means	.30	**.55**	.01
Q19	Identify Ho and Ha for ANOVA	.32	**.48**	.15
Q23	Identify graph of F distribution	.29	**.46**	.14
Q13	Calculate normal distribution probability	.29	**.45**	.16
Q3	Identify experimental units	.32	-.18	**.62**
Q4	Identify factors in an experiment	.28	-.19	**.59**
Q22	Interpret HT vs CI results	.33	.05	**.54**
Q17	Identify significance of a test	.28	.11	**.46**
Q20	Identify Ho and Ha for paired test	.15	-.14	**.43**
Q5	Identify response variable in an experiment	.17	-.05	**.43**
Q9	Identify Ho and Ha for paired test	.28	.20	**.41**
Q10	Identify Type I and Type 2 errors	.15	-.01	**.38**
Q12	Identify appropriate hypothesis test	.16	.04	**.38**
Q1	Identify statistical technique	.30	.30	**.34**
Q8	Identify CI assumptions	.16	.12	**.33**
Q18	Interpret 95% CI for means	.11	.06	**.31**
Q2	Identify the most representative sample	.11	.09	**.29**
Q7	Calculate degree of freedom	.20	.25	**.27**
Eigenvalues			4.28	1.58
Percent of variance			18.60	6.85
Alpha			.72	.69

Note: h^2 = communalities; No. = item number. Factor loading used to identify the components appears in bold.

Next, I conducted CFA of the two-factor model with S2PE scores. The path diagram for this model is shown in Figure 7.

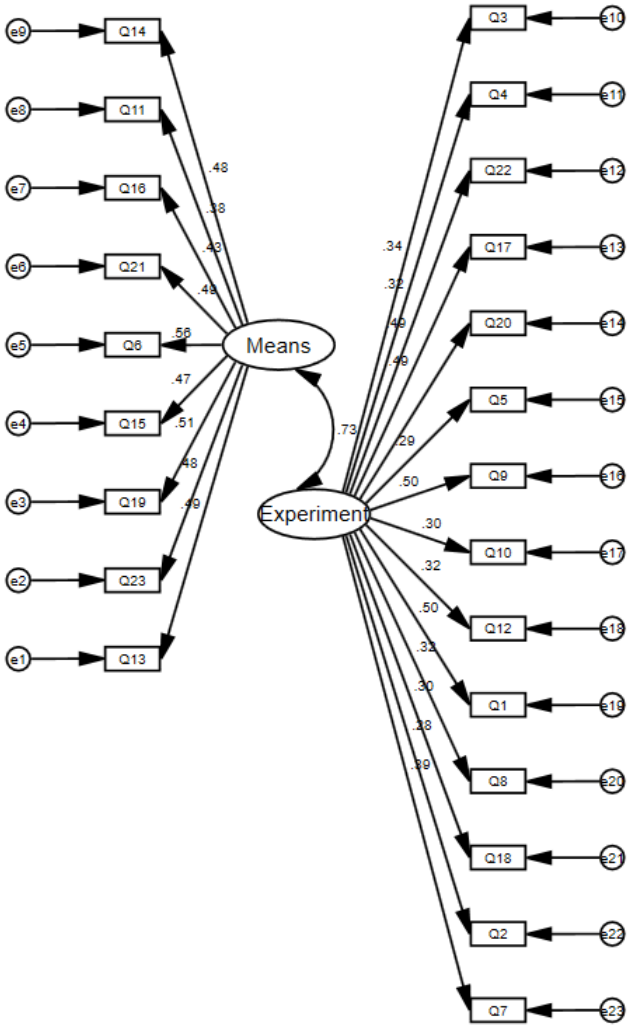


Figure 7. Two-factor model for S2PE.

According to Figure 7, the coefficients in Means ranged from .38 to .56, and the coefficients in Experiment ranged from .28 to .50. The correlation between the Means and Experiment factors was .73. The chi-square test was significant (χ^2 (229) = 568.13, *p* < .05). The ratio χ^2/df was 2.48 > 2, and CFI was .81< .9, indicating weaker model fit than the model fit with S2UPE scores. This difference could be explained by the fact that the number of cases, the number of exams without missing responses, was 191 more on

S2PE than on S2UPE. The value of χ^2 depends on the number of cases: if other parameters are the same, the bigger the number of cases is, the higher the χ^2 value is (Tabachnick & Fidel, 2007). Other indices were very similar to the fit indices of S2UPE model (*RMR = .04, GFI = .93, AFFI = .91, REMSA = 0.4*). To improve a model fit, Tabachnick and Fidell (2007) recommended using the largest modification indices. The two largest indices in S2PE model were 92.60 between e10 and e11 and 50.53 between e13 and e16. In the second factor Experiment, I constructed two covariance pathways between e10 and e11 and e13 and e16. Because both paths were inside of one factor, the internal consistency of the factor was preserved (Tabachnick & Fidell, 2007). The modified path diagram is shown in Figure 8.

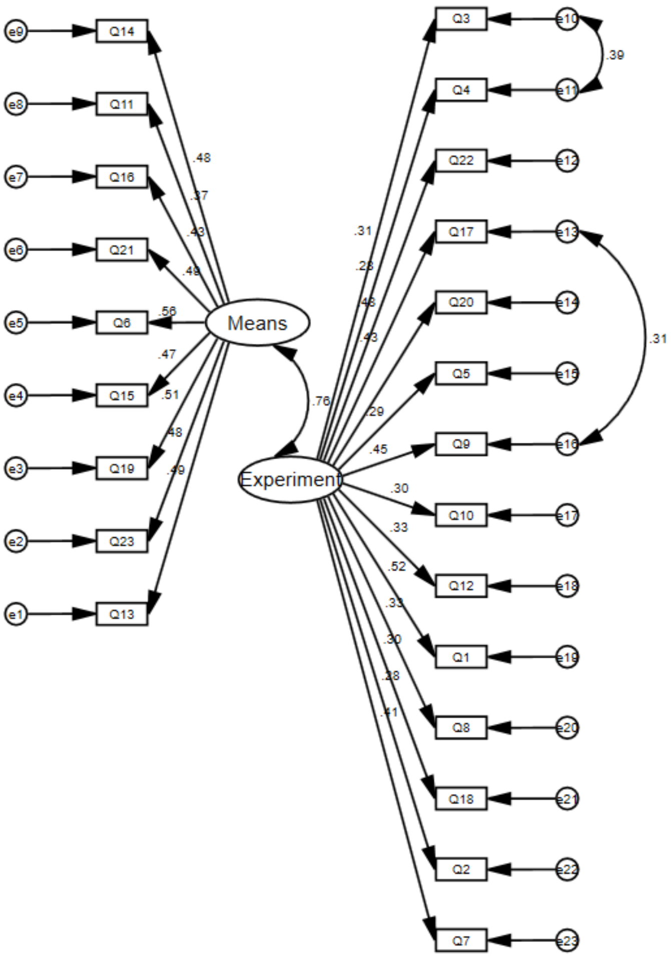


Figure 8. Modified two-factor model for S2UPE.

To test whether the changes improved the model fit, I ran the test for the modified model. The chi-square decreased to 415.7, the ratio χ^2/df became 1.83, and the CF1 increased to .90. The correlation between the factors increased from .73 to .76. The modified model had an adequate fit, suggesting that the construct validity of S2PE was acceptable. The results of the maximum likelihood tests for the initial, modified two-factor model with S2UPE scores, and other models for comparison are provided in Table 44.

Table 44

Results of CFA for Two-factor Model of S2PE Scores and all other Models

Model	*N*	χ^2	*df*	χ^2/df	*GFI*	*AGFI*	*CFI*	*RMR*	*RMSEA*
S1PE	536	401.49	229	1.75	.94	.93	.92	.03	.03
S1UPE	439	339.40	229	1.48	.94	.92	.90	.03	.03
S2UPE	444	402.88	229	1.76	.93	.91	.91	.04	.04
S2PE	653	568.13	229	2.48	.93	.91	.81	.04	.04
Modified S2PE	653	415.70	227	1.83	.95	.93	.90	.04	.04

As seen in Table 44, the model fit indices were similar across all four exams, suggesting equivalency of construct validity across the exams. The number of exams without missing responses was bigger on the proctored exams in both Set 1 and Set 2. Interpretations of these results are given in Chapter 5.

Summary

When equivalent automatically-scored web-based exams with the same security mechanisms were used, there was no significant difference in students' performance on proctored and unproctored exams in Set 1 (M_{S1PE} =67.26, M_{S2PE} = 68.56, *p* = .12). The null hypothesis of the first research question in Set 1 was retained. However, a significant difference in scores on unproctored and proctored exams with the medium effect size was observed in Set 2: the students performed better on the proctored exam, which was administered second (M_{S2UPE} =63.12, M_{S2PE} = 69.51, *p* <.001, η^2 =.15). The null hypothesis of the first research question in Set 2 was rejected.

There was a significant difference in students' performance with respect to the order in which the proctored and unproctored exams were administered. The students' performance was not statistically different on proctored and unproctored exams when the unproctored exam was administered second, and statistically different with, the small effect size, when the unproctored exam was administered first (*p* <.001, η^2 =.08). The null hypothesis of the second research question was rejected.

There was no significant difference in students' performance on equivalent automatically-scored proctored and unproctored web-based exams with respect to the course delivery mode in Set 1 (*p* = .13). Similarly, no significant difference in scores with respect to the course delivery mode was observed in Set 2 (*p* = .92). The null hypothesis of the third research question was retained in both sets.

There was a significant difference with the small effect size in students' performance on proctored and unproctored exams with respect to the instructor of the course in Set 1 (*p* =.001, η^2 =.030). The null hypothesis of the fourth research question in Set 1was rejected. No significant difference in students' scores on unproctored and proctored exams with respect to the instructor of the course was observed in Set 2 (*p* =.980). The null hypothesis of the fourth research question in Set 2 was retained.

According to the additional statistical tests, the interaction order*course delivery mode effect (*p* = .21) and format*course delivery mode effect (*p* = .13) were not significant. All other interaction effects were significant: order*instructor (*p* = .01, η^2 = .04), order*format ($p \leq .004$, η^2 =.12), order*format*mode ($p \leq .004$, η^2 =.02), and order*format*instructor ($p \leq .004$, η^2 =.06). In the group of students who took the second version of the unproctored exams, V2 group, there was no significant difference in scores on proctored and unproctored exams in Set 1 (*p* =.43), but there was a significant difference in performance with a large effect size in Set 2 (*p* <.001, η^2=.310). Additionally, in V2 group, the order effect was significant (*p*=.02, η^2=.04), but no significant order*format interaction was observed (*p* = .89). In the group of students who had extended test time, all effects were not significant (format effect: *p* = .08; order effect: *p* = .69; order*format effect: *p* = .39). The students who took both exams in Set 2 in unproctored format performed significantly better, with a large effect size, on the second exam, which was administered asynchronously, than on the first synchronously administered unproctored exam (*p* =. 004, η^2=.24).

When the scores of the six identical questions were analyzed, there was no significant order effect (*p* =.053), order*format effect (*p* =.177), format*mode effect (*p* =.793) and format*instructor effect (*p* =.213). However, a significant format effect (*p* =.004, η^2=.092), order*mode effect (*p* =.018, η^2=.018), and order*instructor effect (*p* =.018, η^2=.018) were observed.

There was no significant difference in the students' GPA across the course delivery modes (*p* = .474) and across the instructors (*p* = .248). A significant difference with a small effect size in students' age across the course delivery modes was observed (*p* $\leq$.004, η^2=.057): the students in the hybrid sections were significantly older than students in face-to-face and online groups. There was no significant difference in students' age across the instructors (*p* = .056).

A significant decrease in the dropout rate after the proctored and unproctored web-based exams were implemented was observed (*p* <.001, η^2=.12). Reliability analysis of the study’s exams indicated high reliability with alpha coefficients ranged from .76 to .84. According to the PCA and CFA, the construct validity of the four exams was adequate.

I use the findings of the additional statistical tests to interpret the answers to the four research questions in Chapter 5. In this chapter, I also describe how the results of my study resonated with the findings of the literature discussed in Chapter 2. Additionally, Chapter 5 includes interpretations of the study’s results in the context of the theoretical framework. I discuss the limitations of the study, recommendations for further research, and implications for social change at the end of Chapter 5.

Chapter 5: Discussion, Conclusions, and Recommendations

The purpose of this quantitative quasi-experimental study was to investigate whether inconvenient and expensive proctoring is necessary when systematically selected security mechanisms are used. The study was conducted to analyze whether the security mechanisms utilized by the department on web-based exams in an introductory statistics community college course can be an effective alternative to proctoring. To fulfill the study's purpose, I examined the relationship between the format in which the equivalent automatically-scored secured, web-based exams were administered, proctored versus unproctored, and the exam scores. Additionally, the order, course delivery mode, and instructor effects on students' scores on the proctored and unproctored exams were analyzed. To rule out confounding variables relevant to the main analyses, I tested the hypotheses of the research questions with the scores of the six identical questions, reliability and construct validity of the study's exams, attrition bias due to drop out, and compared students' GPA and age across the course delivery modes and instructors. The entire population under investigation included the main group of students, the students who took all exams with all security mechanisms, the students who took the second version of the unproctored exams, V2 group, and the students who had extended test time, Ext. time group.

According to the analysis in the main group, there was no significant difference in students' scores when the proctored exam was administered first and unproctored exam was administered second (p = .12). However, the participants performed significantly better on the proctored exam when the unproctored exam was administered first and the proctored exam was administered second (*p* <.001, η^2 =.15). The order effect was significant (*p* <.001, η^2 =.08). Regardless of the order in which the proctored and unproctored exams were administered, there was no significant difference in students' scores across the course delivery modes: face-to-face, hybrid, online (Set 1: *p* =.13; Set 2: *p* = .92). When the first pair of proctored and unproctored exams was administered, there was a significant difference in students' scores across the instructors (*p* =.001, η^2 =.030). The pairwise test indicated that the difference occurred between Instructor 4 and Instructor 3 (*p* =.008) and Instructor 4 and Instructor 6 (*p* =.002). No significant difference in students' scores with respect to the instructors was observed during the second phase of the exams' administration (*p* =.980).

According to the additional statistical tests, in the main group, the interaction order*course delivery mode effect (*p* = .21) and format*course delivery mode effect (*p* = .13) were not significant, indicating that insignificant order and format effects manifested similarly across the course delivery modes. Like the main group, in the group of students who took the parallel versions of the unproctored exams (V2 group), the format effect was insignificant in Set 1 (*p* =.43) and significant in Set 2 with the large effect size (*p* <.001, η^2 =.310); the order effect was also significant (*p* =.02, η^2 =.04). In contrast, the format effect in both sets and order effect were not significant in the group of students who had extended test time (format effect: *p* = .08; order effect: *p* = .69). The students who took the first exam in Set 2 in synchronous unproctored format and the second exam in asynchronous unproctored format performed significantly better, with the large effect size, on the exam that was administered asynchronously (*p* =. 004, η^2 =.24).

When the scores of the six questions that were identical on all four exams were analyzed, the format effect was significant (*p* =.004, η^2 =.092), but, unlike the main group, the order effect was insignificant (*p* =.053). The interaction order*mode effect with the scores of the six questions was significant, indicating that the insignificant order effect manifested differently across the course delivery modes.

The participant's GPA was not significantly different with respect to the course delivery modes (*p* = .474) and instructors (*p* = .248). The students' age was not statistically different across the instructors (*p* = .056), but a significant difference with the small effect size was observed with respect to the course delivery modes (*p* <.001, η^2 =.057). According to the Bonferroni post hoc test, the hybrid students were significantly older than online and face-to-face students (*ps* <.001).

The dropout rate decreased significantly when the proctored and unproctored exams were implemented (*p* < .001, η^2 =.12). The reliability and factor analyses indicated that the study's exams had adequate reliability and construct validity. During the confirmatory factor analysis, I found that the number of students who respond to all exam question was bigger on the proctored exams (S1PE: *n* = 536; S2PE: *n* = 653) than on the unproctored ones (S1UPE: *n* = 439; S2UPE: *n* = 444).

Interpretations of the Findings

In this section, the results of the conducted analyses for each study's research question are interpreted and compared with the results found in the literature described in Chapter 2. The findings of the additional statistical tests are also compared with related findings in the literature and used for further interpretations. Additionally, the results are analyzed and interpreted in the context of the study’s theoretical framework.

Interpretation of the Results of Testing the Null Hypothesis of RQ1

The first research question and the corresponding hypotheses were stated in the following form:

When equivalent automatically-scored web-based exams with the same security mechanisms are used,

RQ1: Is there a relationship between the exam format (IV1), proctored versus unproctored, and student scores (DV)?

> $H_0$1: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms.
>
> H_A1: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms.

In the group of students who took the exams with all security mechanisms (main group) and the group of the students who took the second version of the unproctored exams due to a schedule conflict (V2 group), there was no significant difference in the students’ scores on the proctored and unproctored exams in Set 1 (main group: p =.12; V2 group: p =.43), but the scores were significantly lower on the unproctored exam in Set

2 (main group: $p < .001, \eta^2 = .15$; V2 group: $p <.001, \eta^2 = .31$). In the group of students who had extended test time (Ext. test time group), there was no significant difference in scores on proctored and unproctored exams in both sets (Set 1: p =.20; Set 2: p =.17). All these findings suggest that if the students tried to cheat during the unproctored exams, they were not successful. Therefore, the security mechanisms used by the department were effective.

The insignificant results of testing the null hypothesis of RQ1 in Set 1 of the main and V2 groups and in both sets of the Ext. test time group parallel the findings of Beck (2014) and Stack (2015). Both researchers found no significant difference in scores on proctored and unproctored exams (Beck: t =.347, p > .05, the exact p was not stated; Stack: b =1.08, p > .05, the exact p was not stated). Beck used randomization, blocked backtracking, restricted test time, and cheating warning statements during unproctored exams in a university economics course. Stack, in addition to the mechanisms used by Beck, administered all unproctored exams in a university criminology course synchronously, emphasizing the importance of synchronous testing. In the given study, the results of the comparison of the scores on the synchronous and asynchronous unproctored exams in the group of students who took both exams in Set 2 in unproctored format (UP group) reinforce the importance of synchronous testing as a security mechanism: The scores on the asynchronous unproctored exam were significantly higher than on the synchronous unproctored exam with the large effect size (p =. 004, η^2 =.24). Moreover, the difference in scores on the two exams in the UP group was about 3% higher than the difference in scores on the corresponding exams in the main group.

The significant lower results on the unproctored exam in Set 2 of the main and V2 groups contradict the findings of Arnold (2016), Fask et al. (2015), and Varble (2014) whose students had significantly higher scores on the unproctored exam than on the proctored one (Arnold: $p < .001$; Fask et al.: $p = .0000028$; Varble: $p < .001$) and parallel the findings of Ladyshewsky (2015), whose students had lower scores on unproctored exams (*p*-value was not stated). Similar to the order of the exams in Set 2 of the given study, in Arnold (2016), Fask et al. (2015) and Varble's (2014) studies, the unproctored exam was followed by the proctored one. However, Arnold used only randomization of items and restricted test time with a cohort of freshmen university students, and Fask et al. did not use any security mechanisms in an introductory university course. While Varble utilized systematic approach to the selection of the security mechanisms in a business university course, he did not incorporate synchronous testing, and many test items on Varble's exams were at remember level of Bloom's taxonomy. Ladyshewsky, whose graduate business students had lower scores on unproctored exams, in addition to mechanisms used by Beck, used higher thinking exam items. The department where the study took place synthesized the best practices accumulated by the previous research and incorporated a carefully selected combination of security mechanisms to reduce academic dishonesty.

The selection of the mechanisms done by the department was based on the taxonomy of cheating prevention techniques (Tinkelman, 2012; Varble, 2014) rooted in the fraud triangle theory (Cressey, 1950). This theoretical framework explains which security mechanisms have the potential to reduce need, rationalization, and opportunity factors needed for cheating to take place. The department utilized synchronous testing, higher order thinking items, randomization, restricted test time, one item per page, blocked backtracking, deferred feedback, and multiple versions of exams to minimize opportunity to cheat. Additionally, the instructors used the cheating warning statement to reduce rationalizatio*n* and created the study guides and web-based practice tests to decrease the need to cheat. The results of testing the null hypothesis of RQ1 confirm that the combination of the security mechanisms selected based on the taxonomy of cheating prevention techniques rooted in the fraud triangle theory was effective. Particularly, the findings of RQ1 analyses in Set 1 in the main and V2 groups and both sets in Ext. test time group confirm that when factors for cheating are minimized, student performance on proctored and unproctored exams may be similar and proctoring may not be necessary.

The fact that in Set 2, in the main and V2 groups, the scores on the unproctored exam were significantly lower than the scores on the proctored one suggest that the used mechanisms did not allow the participants to be successful with cheating. However, it raises a question why in Set 2 the students' performance on the unproctored exam was significantly worse than on the proctored one. To answer this question, I compared the scores of all four exams in all groups.

In the main group, in Set 1, the mean score on the proctored exam was 67.26% and on the unproctored one 68.56%. In Set 2, the mean score on the unproctored exam was 63.82% and on the proctored one 69.51%. Thus, the scores increased by 1.3% between the first (S1PE) and the second (S1UPE) exams, decreased by 4.7% between the second (S1UPE) and third (S2UPE), and increase again by 5.7% between the third

(S2UPE) and fourth (S2PE) exams. As it was discussed previously, the exams within each set were parallel tests that had the same items with different numerical values. The exams between the sets covered slightly different topics, but the equivalency of these exams was established by the faculty, experts in the discipline. The PCA and CFA analyses described in Chapter 4 confirmed that the reliability and construct validity of all four exams were similar. Additionally, the scores on the exams in Set 1 and the proctored exam in Set 2 were close, supporting equivalency of exams between the sets. Moreover, according to the additional analysis, the pattern of the students' scores of the six identical items was similar to the pattern of the scores with all test items: the means went up between the first and second exams, went down between the second and third exam, and then went up again between the third and fourth exams (M_{S1PE} = 0.81, M_{S1UPE} = 0.90, M_{S2UPE} = 0.86, M_{S2PE} = 0.98).

All these facts suggest that in Set 2 the students could perform significantly worse on the unproctored exam than on the proctored one not because of the possible differences in exams, but due to some other factors. The exams within each set were 7-10 days apart, while Set 2 exams were administered in 1 month after Set 1 exams. According to Spitzer (1936) and Falleti et al. (2006), in 30 days, individuals can forget up to 90% of acquired knowledge and skills. Therefore, in the main group, the scores could increase between the first two exams due to learning, decrease between the second and third exam due to forgetting, and increase again between the third and fourth exams due to learning. Another factor of the decrease in scores between the second and third exams could be end of the semester fatigue (Koschel, Young, & Navalta, 2017), which the students could overcome on the last exam.

In V2 group, the group of students who took the second version of the unproctored exam, the pattern of the scores in Set 1 was slightly different from the pattern of the scores observed in the main group. In Set 1, the mean score on the proctored exam was 63.97% and on the unproctored one 61.12%. In Set 2, the mean score on the unproctored exam was 55.94% and on the proctored one 65.83%. Thus, the scores decreased by 2.9% between the first and the second exams, decreased further by 5.2% between the second and third exams, and increased by 9.9% between the third and fourth exams. The fact that the scores on the unproctored exams were lower than the scores on the proctored exams indicates that the security mechanisms in V2 group were effective. However, the question is why the students in V2 performed worse on both unproctored exams than on the proctored exams.

The participants in V2 group could be sensitive to the environment where the unproctored exam took place. Fask et al. (2014) found that students can have lower scores on an unproctored exam than on the equivalent proctored exam due to possible distractions at home. During the reliability and factor analyses, I found that the number of exams with missing responses was larger on the unproctored exams than on the proctored ones (S1PE: *n* = 536, S1UPE: *n* = 439; S2UPE: *n* = 444; S2PE: *n* = 653), which could also indicate that, for some students, the home environment was less productive than class environment. Also, the facts that the students in V2 group rescheduled the unproctored exams, unlike the main group, earned lower scores on the unproctored exam than on proctored one in Set 1, and had lower scores on all four exams than the students in the main group may suggest that the V 2 group students could be busier than the students in the main group. This explanation is based on Ladyshewsky' (2015) observations and can be explore in future research.

In the Ext. test time group, in Set 1, the mean score on the proctored exam was 57.43% and on the unproctored one 63.79%. In Set 2, the mean score on the unproctored exam was 62.77% and on the proctored one 68.71%. Thus, the scores increased by 6.4% between the first and the second exams, slightly decreased by 1% between the second and third exams, and increase by 5.9% between the third and fourth exams. The fact that the scores on the unproctored exam in Set 2 were lower than on the proctored one suggests that the security mechanisms utilized by the department were effective in the Ext. test time group as well.

Overall, the patterns in scores of the Ext. test time group and the main group were similar: the scores went up, then down, then up again. However, the difference in scores between the second and third exams was much smaller in the Ext. test time group than the corresponding difference in the main group (Main group: diff = 4.9%; Ext. test time: diff =1.0%). These findings may indicate that it could take more time for the students in the Ext. test time group to learn the material and acquire web-based test-taking skills than the students in the main group, but by the forth exams the scores of the students with extended time increased to the level of the students in the main group. This interpretation is based on the findings of Lee et al. (2010) who examined the web-testing experience of students with special needs and found that when this population of students acquire web-based test-taking skills, their feel comfortable in a web-based environment and their performance increases.

In summary, the results of RQ1 analyses done with the scores of community college students confirm Beck (2014) and Stack’s (2015) findings obtained with the scores of university students that systematically selected security mechanism lead to no significant difference in students’ performance on web-based proctored and unproctored exams. Additionally, the RQ1 results reinforce and expend Varble’s (2014) idea that security mechanisms might be especially effective if they are selected based on the taxonomy of cheating prevention techniques rooted in the fraud triangle theory. With a carefully selected combination of security mechanisms, students’ performance on proctored and unproctored exams may be similar and proctoring might not be necessary.

Interpretation of the Results of Testing the Null Hypothesis of RQ2

The second research question and the corresponding hypotheses were stated in the following form:

When equivalent automatically-scored web-based exams with the same security mechanisms are used,

RQ2: Is there a relationship between the order (IV2) in which proctored and unproctored exams are administered and student scores (DV)?

> $H_0$2: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the order in which exams are administered.

> H_A2: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the order in which exams are administered.

In the main group of students who took all four exams, there was a significant difference in the students' scores on the proctored and unproctored exams with respect to the order in which the exams were administered ($p < .004$, $\eta^2 = .08$). In this group, the difference in scores when the unproctored exam followed the proctored one was 0.1%, while the difference in scores when the proctored exam followed the unproctored one was 6.0% ($M_{S1PE} = 71.73$, $M_{S1UPE} = 71.86$, $M_{S2UPE} = 64.25$, $M_{S2PE} = 70.22$). A medium significant order*format interaction effect (p < .001, $\eta^2 = .13$) suggested that the found significant order effect manifested differently across the formats. These findings may be explained as follows.

As it was described in the first three chapters, the interval between the exams within each set was 7-10 days, between the sets 30 days. The students studied hard for the first exam and remembered the material well to earn almost the same score on the second exam. To perform at the same level on the third exam, the participants had to learn several new concepts and review some previous topics. It is possible that the students did not realize that in 30 days they forgot the previous material and did not review the concepts well, which could result in a lower score on the third exam. After the third exam, the participants understood that they forgot the previous material and reviewed the material better, which resulted in the increase in scores to the level comparable with the scores of the first two exams. This explanation is based on research on effect of learning, according to which student success may depend on appropriate learning strategies and self-regulation (Roediger & Karpicke, 2006; Tuckman, 2003). Another explanation of the lower scores on the third exam may be related to the end of the semester fatigue.

The students could study hard the new material and reviewed the previous material well. However, due to the end of the semester fatigue, they could not perform on the unproctored exam in Set 2 at the highest of their potential. For this reason, the scores on the exam were lower. Understanding the importance of the final exam, the students mobilized all their potential and earned the scores comparable with their previous performance. This interpretation is rooted in findings of Koschel et al. (2017) who studied end of the semester fatigue in college students.

The results of the RQ2 analysis in the main group contradict the findings of Templer and Lange (2008) who found insignificant order effect on proctored and unproctored future employee's personality test in their laboratory experiment with university students (p > .05, the exact *p* was not stated). The difference in the results can be attributed to the fact that Templer and Lange utilized exactly the same test on all exams, while in the given investigation the exams between the sets were on slightly different topics. The interpretation of the results of testing the null hypothesis of RQ2 with the scores on the six identical questions that follows supports this idea.

In the main group, the order effect with the scores of the six identical questions was not significant ($p = .053$). In spite of the fact that the mean scores of the six questions on both exams in Set 1 were lower than the mean scores in Set 2 ($M_{S1PE} = 0.81$, M_{S1UPE}

= 0.90, M_{S2UPE} = 0.86, M_{S2PE} = 0.98), the differences in the scores on the proctored and unproctored exams within each set were almost the same, about 0.1 of a point. The order*format interaction effect with the scores of the six questions was not significant (p = .18). Thus, regardless of whether the exam was proctored or unproctored, in the main group, the scores of the six questions were higher on the exam that was administered second by the same number of points. These findings, based on the six identical test items, suggest that when proctored and unproctored exams include identical items, the order effect is not significant and an increase in scores occurs due to learning. Templer and Lange (2008) obtained similar results with the same future employee's personality test administered in proctored and unproctored formats. The researchers found that the order in which the test was administered did not influence the participants' scores and attributed an increase in scores on some batteries of the test due to practice effect.

In V2 group of students who took all four exams, a large significant order effect was observed (p =.02, η^2 = .38). The scores decreased by 6% when the unproctored exam was administered second, and increased by 7% when the unproctored exam was administered first (M_{S1PE} = 67.24, M_{S1UPE} = 61.26, M_{S2UPE} = 60.18, M_{S2PE} = 67.21). In this group, the order*format interaction effect was not significant (p =.89). Regardless of the order in which the proctored and unproctored exams were administered, the almost identical scores on the proctored exams were higher than also almost identical scores on the unproctored exams. As I described in the interpretation of the RQ1 analysis using the findings of Fask et al. (2014), the students in V2 could have lower scores on both unproctored exams due to possible distractions in the unproctored environment. The scores of these students on all exams were lower than the corresponding scores of the students in the main group, which could be explained by the fact that the students in V2 may have been busier than the students in the main group. Ladyshewsky (2015) observed that busier students tend to have lower scores.

In Ext. time group of students who took all four exams, the order effect was not significant (*p* = .69). The difference in scores when the unproctored exam followed the proctored one was 10.3%, and the difference in scores when the proctored exam followed the unproctored one was 7.1%. The order*format interaction effect was also not significant (*p* = .39). Regardless of whether the exam was proctored or unproctored, the scores of the students with extended test time were higher on the exam that was administered second most likely due to learning. Because there were only 10 students who took all four exams with extended test time, the generalizability of this results might be limited.

In summary, to the best of my knowledge, the given study is the first study in which the order effect was examined in a natural educational setting. The insignificant result in the order effect analysis with the scores of the six identical questions confirms the findings of Templer and Lange (2008) who found no significant order effect in their laboratory experiment with university students' scores on the same future employee's personality test administered in proctored and unproctored format. The significant result in the order effect analysis with the scores of all items contradicts the findings of Temple and Lange, suggesting that there might be a significant difference in scores on proctored and unproctored exams with respect to the order in which the exams are administered if exams cover slightly different topics. The fact that the scores in the main group and on the six identical items were lower on the unproctored exams regardless of the order in which the exams were administered indicates that the students were not successful with cheatings in both sets even if they tried to cheat.

Interpretation of the Results of Testing the Null Hypothesis of RQ3

The interpretations of the results of RQ3 analyses are limited to the main group because there were none or just a few students in V2 and Ext. time groups in the hybrid and online sections. The third research question and the corresponding hypotheses were stated in the following form:

When equivalent automatically-scored web-based exams with the same security mechanisms are used,

RQ3: Is there a relationship between the course delivery mode (IV3), (a) web-assisted face-to-face, (b) hybrid, (c) fully online, and students’ scores (DV)?

> $H_0$3: There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the course delivery mode.
>
> H_A3: There is a significant difference in students' performance on automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the course delivery mode.

In the main group, in both Set 1 and Set 2, there was no significant difference in the participants' scores on proctored and unproctored exams with respect to the course delivery mode (Set 1: *p* = .13; Set 2: *p* =.92). These results parallel the findings of Beck (2014) who found no significant difference in scores on proctored and unproctored exams in face-to-face (*N* = 60), hybrid (*N* = 21), and online (*N* = 19) sections of a university economics course. Beck had small sample sizes in the hybrid and online groups. The results of the given study with bigger sample sizes (face-to-face: *N* = 642; hybrid: *N* = 91; online: *N* = 55) confirm and expand Beck's findings. The format*mode interaction effect was not significant in Set 2 (*p* =.83), but significant in Set 1 (*p* <.001), indicating that in Set 1 the insignificant format effect manifested differently across the course delivery modes.

In Set 1, the scores of students in face-to-face and online sections were about 1% higher on the unproctored exam than on the proctored one, while the scores of the students in the hybrid sections were about 5% higher on the proctored exam than on the unproctored one. These results can be explained by the fact that the hybrid students were about four years older than face-to-face and online students, and this difference in age was significant (*p* <.001). No significant difference was found in GPA across the course delivery modes, indicating that the observed difference in scores could not be attributed due to different academic abilities. Older students could have more distractions at home during unproctored exams than younger students because of their busier life, kids and other family obligations. These explanations are rooted in the following combination of findings of Ladyshewsky (2015) and Fask et al. (2014). Ladyshewsky found that the hybrid students whose scores were lower on the unproctored exams in a graduate business course were older than the students in the face-to-face sections of the same course whose scores were lower on the proctored exams. Ladyshewsky explained this difference in scores by busier life of older students. Fask et al. found that lower scores on the unproctored exams could be attributed to distractions at home or other out of class environment. Thus, the findings related to the significant format*mode interaction effect of the given study suggest a possible relation between significantly older age, lower scores on the unproctored exams, and distractions in an unproctored environment. A possible further exploration of this relationship is discussed in the Recommendation section of this chapter.

In summary, the insignificant course delivery mode effect found in the RQ3 analyses with the scores of introductory statistics community college students confirm and expand the findings of Beck (2014) who found no significant difference in students' scores on proctored and unproctored university economics exams in a sample of 60 face-to-face students and small samples of 21 hybrid and 19 online students. The significant format*mode effect identified in the RQ3 analyses expands Ladyshewsky's (2015) results found with older graduate business students and Fask's et al. (2014) findings obtained with undergraduate statistics students. The significant*mode effect suggests a possible relationship between older age, lower scores on the unproctored exams, and distractions in an unproctored environment.

Interpretation of the Results of Testing the Null Hypothesis of RQ4

The interpretations of the results of RQ4 analyses are limited to the main group because some instructors did not have students who took the second version of the unproctored exams (V2 group) or students with extended test time (Ext. time group). The fourth research question and the corresponding hypotheses were stated in the following form:

When equivalent automatically-scored web-based exams with the same security mechanisms are used,

RQ4: Is there a relationship between the instructor (IV4) and students' scores (DV)?

> $H_0$4**:** There is no significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the instructor of the course.
>
> H_A4: There is a significant difference in students' performance on equivalent automatically-scored unproctored and automatically-scored proctored introductory statistics web-based exams with the same security mechanisms with respect to the instructor of the course.

In the main group, the group of students who took all exams with all security mechanisms, there was a small significant instructor effect in Set 1 (p =.001, η^2 =.030). According to the pairwise comparison described in Chapter 4, the differences in the exams' scores in Set 1 were significantly different between Instructor 4 and Instructor 3 groups (p =.008) and Instructor 4 and Instructor 6 groups (p =.002). These significant differences between the instructors' groups can be explained by the fact that Instructor 4 had no experience in administering online exams at the time when the department incorporated the web-based testing, while Instructor 3 had been administering online exams for five semesters and Instructor 6 for three semesters. These results parallel the findings of Ladyshewsky's (2015) who found different patterns with the scores of proctored and unproctored exams of the graduate business students taught by less experienced and more experienced instructors. However, Ladyshewsky did not test these differences statistically and discussed the instructors' experiences in general, not specifying their experiences with administering online tests. Suggestions on further research about the relationship between instructors' experience with online testing and students' score on proctored and unproctored exams are discussed in Recommendations section of this chapter.

In Set 2, the instructor effect was not significant (p = .98), indicating that there was no significant difference in students' scores on proctored and unproctored exams with respect to the instructors. This result may indicate that Instructor 4 became more comfortable with administering online tests by the end of the semester to the level at which no significant instructor effect was observed. This explanation parallels the findings of Hanson and Robson (2004) who found that with some training instructors feel more comfortable with administering online tests.

The insignificant format*instructor interaction effect was observed in Set 1 (p = .08), indicating that in this set the significant format effect manifested similarly across the instructors. However, in Set 2, the format*instructor interaction effect was significant with a small effect size (p <.001. η^2 =.05): the scores in Instructor 1 and Instructors 2 groups increased from unproctored to proctored exam by less than 1%, while the scores in all other instructors increased by more than 5% (I$_1$: M_{S2UPE} = 67.58, M_{S2PE} = 68.35; I$_2$: M_{S2UPE} = 65.53, M_{S2PE} = 66.21; all other Is: average of M_{S2UPE} = 61.78, average of M_{S2PE} = 70.62). This significant difference could be explained by the fact that although Instructor 1 and Instructor 2 had some previous experience with online testing, they joined the team of statistics instructors right before the implementation. These findings also support a possible relationship between instructors' experiences in web-based teaching and students' scores on proctored and unproctored exams discussed by Ladyshewsky (2015).

In summary, the significant instructor effect obtained during RQ4 analyses with the scores of community college students parallel the observations of graduate students' scores on proctored and unproctored exams with respect to instructors done by Ladyshewsky's (2015). The significant instructor effect fills the gap by statistically confirming Ladyshewsky's hypothesis that the differences in students' scores on proctored and unproctored exams with respect to instructors can occur due to distinctions in instructors' experiences. The found significant format*instruction interaction effect gives another evidence that supports the existence of the relationship between instructors' experience and students' performance on proctored and unproctored exams.

Limitations of the Study

As discussed in Chapter 1, the major limitation of the study was its quasi-experimental nature. There was no random assignment into the groups with respect to the course delivery mode and instructor. However, the results of high-quality quasi experiments conducted in natural educational settings might be more valuable than hardly applicable true experiments with random assignments into groups because randomization is almost never possible in a regular educational practice (Kim & Steiner, 2016). To determine the degree to which the nonrandom assignment into the groups could influence the trustworthiness of the results, I conducted the additional analysis and found no significant difference in students' GPA across either course delivery modes or instructors and no significant difference in students' age in face-to-face and online groups. The finding that the students were significantly older in the hybrid group than in the face-to-face and online groups was taken into consideration in the Interpretation section of the chapter.

The use of archival data of the department where the study was conducted is another limitation. I had no control of how the data were collected and recorded. However, there were no missing data in the provided spreadsheet. The quality of the recorded information seemed adequate.

Possible high attrition due to dropout during the implementation of the web-based proctored and unproctored exams is listed in Chapter 1 as another limitation. I conducted the additional analysis to compare the attrition rate due to dropout before and after the implementation. According to the additional analysis, the dropout decreased by almost 10% after the web-based exams were implemented.

The entire population under investigation was constituted of the students who took the study's exams with all security mechanisms (the main group), the students who took the second version of the unproctored exams (V2 group), and the students who had extended test time (Ext. time group). The number of students who took all four exams with extended test time was 10. This small sample size limits generalization of the results related to this group of students.

I am the researcher and one of the instructors whose students' scores were collected by the department, which could bring a bias to the study. This bias was neutralized by the involvement of other six instructors and the use of archival data. The archived data were collected by the department as part of regular institutional practice and used originally for program evaluation. The statistics coordinator provided the recoded data spreadsheet with all identifying information removed to protect the identities of the students and instructors.

Recommendations

This study was conducted with community college students at a medium size suburban college in California. A replication of the study at other colleges and universities with different population of students may increase generalizability of the results.

In the design of the given study, the time interval between Set 1 and Set 2 exam administration was 30 days. According to the results of the RQ1 analyses in the main group of students, the students who took the four exams with the same security mechanisms, the time interval of 30 days could lead to lower scores on the unproctored exam in Set 2 due to forgetting. Decreasing the time interval between Set 1 and Set 2 exams from 30 to 7-10 days in future research will allow for testing whether forgetting is the influential factor.

According to the results of the RQ1 analyses in V2 group, the group of students who took the second version of the unproctored exams due to a schedule conflict, the scores on both unproctored exams were lower than on the proctored ones. Based on these results, Fask's et al. (2014) findings about lower scores on unproctored exams due to distractions in unsupervised environment, and Ladyshewsky's (2015) observations of lower scores on unproctored exams of busy students, I concluded that there might be a relationship between lower scores on unproctored exams than on proctored, distractions in an unproctored environment, and business of the participants. This relationship can be studied in future research.

During the RQ1 and RQ2 analyses in the Ext. time group, the group of students who had extended test time, I found no significant difference in the student's scores on proctored and unproctored exams and no significant order effect. However, the sample size of this group was small. Replication of the study with a larger number of students with extended test time in future research will increase generalizability, validity, and reliability of the results.

According to the results of RQ2 analysis there was a significant order effect in the main group of students, but insignificant order effect with the scores of the same students on the six identical questions. I concluded that the significant order effect found with the scores of all items could be attributed to the fact the students did not review well the previous material or experienced end of the semester fatigue effect and could not perform at the highest of their potential. Future research can test whether my conclusion is valid by keeping all other aspects of the quasi-experiment the same, but administering parallel versions of the same exam in both Set 1 and Set 2.

According to the results of the RQ3 analysis, the format*mode interaction effect was significant: unlike face-to-face and online students, the hybrid students had lower scores on the unproctored exams. The additional analysis of the participants' age revealed that the hybrid students were significantly older than the face-to-face and online students. All these facts and observations of Ladyshewsky (2015) and findings of Fask et al. (2014) suggest a possible relationship between older age, lower scores on the unproctored exams, and distractions in an unproctored environment. This relationship can be explored further in future research.

According to the results of the RQ4 analysis, the significant instructor effect and significant format*interaction effect were observed. Based on these findings and Ladyshewsky (2015) observations, the significant results could occur due to differences in the instructors' experience in web-based teaching and online exam administrations. The relationship between instructors' experiences in web-based teaching and assessment and students' scores on proctored and unproctored exams can be studied further in future research.

During the reliability and factor analyses, I found that the number of students who did not respond to at least one test item was smaller on the proctored exams than on the unproctored exams. This fact may confirm Fask et al. (2014) findings that some students can be less productive in an unproctored environment. The relationship between the number of missing responses to test questions, scores on unproctored exams, and an unproductive unproctored environment can be studied in future research.

When I tested the construct validity of the study's exams by using principal factor analysis (PCA) and confirmatory factor analysis (CFA), I found that the two-factor models manifested similarly with the scores of all four exams and had almost identical model fit indices. These findings suggest that regardless of whether the test is administered in proctored and unproctored format, the results of the PCA and CFA of the parallel and equivalent exams should be similar. This relationship between the format of exams and their construct validity can be studied in future research.

During data screening process, I found that about 50 students took the first exam in Set 2 in synchronous unproctored format and the second exam in Set 2 in asynchronous unproctored format. I used this unexpected outcome, which occurred in a natural education setting, to test the format effect and found that the students' scores on the asynchronous unproctored exams were significantly higher than their scores on the synchronous unproctored exams. The relationship between asynchronous and synchronous administration of unproctored exams and students' scores can be studied in future carefully designed studies.

During the additional test on attrition, I found that, after the web-based exams were implemented, the attrition rate due to dropout decreased by almost 10%. According to Seidman (2005) and Tinto (2012), a decrease in the dropout rate leads to an increase of the number of students who successfully complete the course. Whether the success rate in the introductory statistics courses increased after the web-based exams were implemented can be investigated in future research.

Implications

According to the study's findings, a carefully selected combination of security mechanisms based on the taxonomy of cheating prevention techniques can be an effective alternative to expensive and inconvenient proctoring. With the use of this combination of security mechanism, educational institutions will be able to maintain the credibility of their web-based exams administered in unproctored format while providing the students convenience they need. Technology in the form of web-based exams will become an attribute of every day practice, allowing for not spending valuable in-class time on assessment, but rather on learning and instruction. Administrators will be able to use web-based testing for student learning outcome assessment at departmental and institutional levels. Instructors will get credible, convenient, and quick ways to assess their students by utilizing automatically-graded web-based exams. More students with full-time jobs and family commitments will be able to get degrees; more instructors will be willing to teach fully-online courses. The entire society will gain more college graduates with a high potential of becoming valuable professionals in their fields.

The results of the given study empirically confirm that the taxonomy of cheating prevention techniques rooted in the fraud triangle theory is an adequate theoretical framework for a systematic selection of effective security mechanisms. The detailed description of the departmental implementation of the web-based exams into the curriculum of a traditional community college course adds to the body of the best practices of secured online assessment. The results of the analyses of the research questions with the entire population of students under investigation whose age ranged from 14 to 50, including students with extremely busy schedules and students with special needs can be used not only by researcher, but by educators and administrators to utilize web-based exams with any of the subgroups described in the study.

The use of one-group sequential design, controlling for grading and instruction effects, conducting adequate and powerful main and additional statistical tests allowed for a high quality quasi-experiment. This design can be recommended for similar research studies. All literature on principal component analysis (PCA) and confirmatory factor analysis (CFA) I have encountered was mostly on testing psychometric characteristics of Likert-type scales; a few were on assessing qualities of high-stake tests. I applied the methodology used by these researchers to test the construct validity of the web-based introductory statistics exams created by classroom instructors. The PCA and CFA utilized by me can be used by other researchers for testing construct validity of math, physics, or other subject tests composed by teachers.

Conclusion

This quantitative study was conducted to determine whether a combination of security mechanisms systematically selected based on the taxonomy of cheating prevention techniques rooted in the fraud triangle theory can be an alternative of expensive and inconvenient proctoring. The relationship between the format, proctored versus unproctored, of the equivalent automatically-scored secured, web-based exams in the introductory statistics community college course and exam scores was examined.

Moreover, the order, course delivery mode, and instructor effects on students' scores on proctored and unproctored exams were analyzed. To the best of my knowledge, this is the first study with community college students in which the scores on proctored and unproctored web-based exams, the course delivery mode effect, and the instructor effect were investigated, and the first study in which the order effect was examined in a natural educational setting.

The era of high-stake and classroom web-based assessment has begun. Proctored and unproctored web-based exams are in high demand among students and instructors. With the use of security mechanisms carefully selected based on the taxonomy of cheating prevention techniques rooted in the fraud triangle theory, the credibility of unproctored web-based exams can be comparable with the credibility of proctored exams. The integrity of online tests can be maintained without expensive and inconvenient proctoring!

References

Ajzen, I. (2002). Perceived behavioral control, Self-Efficacy, locus of control, and the theory of planned Behavior. *Journal of Applied Social Psychology, 32*(4), 665-683. doi:10.1111/j.1559-1816.2002.tb00236.x

Allen, I. E., & Seaman, J. (2015). Grade change: Tracking online education in the United States with commentary: IPEDS as the new data source. Retrieved from https://www.onlinelearningsurvey.com/reports/gradelevel.pdf

Allen, I. E., Seaman, J., Poulin, R., & Straut, T. T. (2016). Online report card: Tracking online education in the United States. Retrieved from http://onlinelearningsurvey.com/reports/onlinereportcard.pdf

Alt, D., & Geiger, B. (2012). Goal orientations and tendency to neutralize academic cheating: An ecological perspective. *Psychological Studies*, *57*(4), 404-416. doi:10.1007/s12646-012-0161-y

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist*, *36*(10), 1086-1094. doi:10.1037%2F0003-066X.36.10.1086

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.

Anderson, C., & Gades, P. (2017, June). *Proctoring Exams in an Online Environment*. Paper presented at Innovate! Teaching with Technology Conference, Morris, MN.

Ardid, M., Gomez-Tejedor, J., Meseguer-Duenas, J., Riera, J., & Vidaurre, A. (2015). Online exams for blended assessment. Study of different application methodologies. *Computers & Education*, *81*, 296-303. doi:10.1016/j.compedu.2014.10.010

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

Atoum, Y., Chen, L., Liu, A. X., Hsu, S. D., & Liu, X. (2017). Automated online exam proctoring. *IEEE Transition on Multimedia*, 1-15. doi:10.1109/TMM.2017.2656064

Bain, L. Z. (2015). How students use technology to cheat and what faculty can do about it. *Information Systems Education Journal*, *13*(5), 92-99. Retrieved from http://isedj.org/2015-13/n5/ISEDJv13n5p92.html

Bandyopadhyay, K., Barnes, C., & Bandyopadhyay, S. (2015). An investigation of the factors that influence the use of proctoring in online courses. In Qiang Fei (Ed.), *Proceedings of the Association of Collegiate Marketing Educators* (pp.49-50). Huston, TX: Prairie View A&M University. Retrieved from http://acme-fbd.org/wp-content/uploads/2014/04/ACME_2015_Proceedings.pdf#page=58

Bandyopadhyay, K., Barnes, C. (2014). Maintaining academic honesty in online courses. Refereed research paper. Retrieved from www.researchgate.net/publication/264000710_MAINTAINING_ACADEMIC_HONESTY_IN_ONLINE_COURSES on November 27, 2015.

Barnes, C., Paris, B. L. (2013). An analysis of academic integrity techniques used in online courses at a southern university. *Northwest Decision Sciences Institute Annual Meeting Proceedings*. Retrieved from https://www.researchgate.net/profile/Cynthia_Barnes3/publication/264000798_AN_ANALYSIS_OF_ACADEMIC_INTEGRITY_TECHNIQUES_USED_IN_ONLINE_COURSES_AT_A_SOUTHERN_UNIVERSITY/links/00b4953c7d68919e06000000.pdf

Bayazit, A., & Aşkar, P. (2012). Performance and duration differences between online and paper-pencil tests. *Asia Pacific Education Review*, *13*(2), 219-226. doi:10.1007/s12564-011-9190-9

Bayes, A. (2014). Constructing, refining and validating a task for developing reasoning on stabilized frequency distributions in the context of informal inferences. *ICOTS, 9*, 1-6. Retrieved from http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_2E2_SERRADOBAYES.pdf

Beck, V. (2014). Testing a model to predict online cheating -much ado about nothing. *Active Learning in Higher Education*, *15*(1), 65-75. doi:10.1177/1469787413514646

Becker, D., Connolly, J., Lentz, P., & Morrison, J. (2006). Using the business fraud triangle to predict academic dishonesty among business students. *Academy of Educational Leadership Journal*, *10*(1), 37-54. Retrieved from https://search.proquest.com/openview/3d7dae5af62b010d3f002d2de4dc4841/1?pq-origsite=gscholar&cbl=38741

Bedford, W., Gregg, J., & Clinton, S. (2009). Implementing technology to prevent online cheating: A case study at a small southern regional university. *Journal of Online*

Learning and Teaching, *5*(2), 230. Retrieved from http://jolt.merlot.org/vol5no2/gregg_0609.pdf

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*(3), 339-352. doi:10.1076/jcen.20.3.339.822

Bennett, R. (2008). Technology for large scale assessment. *ETS Research Memorandum.* Retrieved from https://www.ets.org/research/policy_research_reports/publications/report/2008

Blazer, C. (2010). Computer-based assessments. Information capsule. *Research Services, Miami-Dade County Public Schools*, *9(18)*, 1-18. Retrieved from https://files.eric.ed.gov/fulltext/ED544707.pdf

Bloom, B. S. (1964). *Taxonomy of educational objectives*. New York, NY: Longmans, Green.

Bolin, A. U. (2004). Self-control, perceived opportunity, and attitudes as predictors of academic dishonesty. *The Journal of Psychology*, *138*(2), 101-114. doi:10.3200/JRLP.138.2.101-114

Brallier, S., & Palm, L. (2015). Proctored and unproctored test performance. *International Journal of Teaching and Learning in Higher Education*, *27*(2), 221-226. Retrieved from https://files.eric.ed.gov/fulltext/EJ1082856.pdf

Brothen, T., & Peterson, G. (2012). Online exam cheating: A natural experiment. *International Journal of Instructional Technology and Distance Learning*, *9*(2), 15-20. Retrieved from http://itdl.org/Journal/Feb_12/Feb_12.pdf#page=19

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The four generations of computerized educational measurement. *ETS Research Report Series*, *1988* (1), 1-148. doi:10.1002/j.2330-8516.1988.tb00291.x

Burke, M., & Bristor, J. (2017). Academic integrity policies: Has your institution implemented an effective policy? *The Accounting Educators' Journal*, *26*, 1-10. Retrieved from http://www.aejournal.com/ojs/index.php/aej/article/view/338

Campbell, D., Stanley J. (1963). *Experimental and quasi-experimental designs for research.* Boston, MA: Houghton Mifflin Company.

Carstairs, J., & Myors, B. (2009). Internet testing: A natural experiment reveals test score inflation on a high-stakes, unproctored cognitive test. *Computers in Human Behavior*, *25*(3), 738-742. doi:10.1016/j.chb.2009.01.011

Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education*, *61*(7), 613-615. doi:10.1021/ed061p613

Catron, D. W. (1978). Immediate test-retest changes in WAIS scores among college males. *Psychological Reports*, *43*(1), 279-290. doi:10.2466/pr0.1978.43.1.279

Catron, D. W., & Thompson, C. C. (1979). Test-retest gains in WAIS scores after four retest intervals. *Journal of Clinical Psychology*, *35*(2), 352-357. doi:10.1002/1097-4679(197904)35:2<352

California Community Colleges Chancellor Office (CCCCO). (2013). *Distance education report*. Retrieved from http://californiacommunitycolleges.cccco.edu/Portals/0/reportsTB/REPORT_DistanceEducation2013_090313.pdf

Champlain, A. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44,* 109-117. doi:10.1111/j.1365-2923.2009.03425.x

Chapman, K. J., Davis, R., Toy, D., & Wright, L. (2004). Academic integrity in the business school environment: I'll get by with a little help from my friends. *Journal of Marketing Education*, *26*(3), 236-249. doi:10.1177/0273475304268779

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1-8. doi:10.1016/j.jebo.2011.08.009

Chen, R. (2014). Teacher's Strategies for Improving Online Learning or Exam Credibility. *Information Technology Journal*, *13*(17), 2674–2681. doi:10.3923/itj.2014.2674.2681

Cheung, H. Y., Wu, J., & Huang, Y. (2016). Why do Chinese students cheat? Initial findings based on the self-reports of high school students in China. *The Australian Educational Researcher*, *43*(2), 245-271. doi:10.1007/s13384-016-0201-z

Choo, F., & Tan, K. (2008). The effect of fraud triangle factors on students' cheating behaviors. *Advances in Accounting Education*, *9*, 205-220. doi:10.1016/S1085-4622(08)09009-3

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593-602. doi:10.1111/1467-8535.00294

Cluskey Jr, G. R., Ehlen, C. R., & Raiborn, M. H. (2011). Thwarting online exam cheating without proctor supervision. *Journal of Academic and Business Ethics*, *4*, 1-7. Retrieved from http://www.aabri.com/manuscripts/11775.pdf

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213. doi:10.1037/h0026256

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155-159. doi:10.1037/0033-2909.112.1.155

College Board. (2016). *Accuplacer*. Retrieved from https://accuplacer.collegeboard.org/professionals/frequently-asked-question

Cook, T., Campbell, D. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Boston, MA: Houghton Mifflin Company.

Corrigan-Gibbs, H., Gupta, N., Northcutt, C., Cutrell, E., & Thies, W. (2015). Deterring cheating in online environments. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *22*(6), 28. doi:10.1145/2810239

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

Creswell, J. W. (2013). *Research design: Qualitative, Quantitative, and Mixed methods Approaches*. Thousand Oaks, CA: Sage publications.

Cummings, R. (2003). Equivalent assessment: achievable reality or pipedream. In *ATN Education and Assessment Conference.* Retrieved from http://w3.unisa.edu.au/evaluations/Full-papers/CummingsFull.doc

Dahalan, H. M., & Hussain, R. M. R. (2010). Development of web-based assessment in teaching and learning management system (e-ATLMS). *Procedia-Social and Behavioral Sciences*, *9*, 244-248. doi:10.1016/j.sbspro.2010.12.144

Dahlstrom, E., Brooks, C., and Bichsel, J. (2014). *The Current Ecosystem of Learning Management Systems in Higher Education: Student, Faculty, and IT Perspectives*. Research report. Louisville, CO: ECAR. Retrieved from https://library.educause.edu/~/media/files/library/2014/9/ers1414-pdf.pdf

Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology*, *31*(3), 207-208. doi:10.1207/s15328023top3103_6

Darwazeh, A., & Branch, R. (2015, November). A revision to the revised Bloom's taxonomy. In M. Simonson, *38th Annual Proceedings*. Paper presented at 2015

Association for Educational Communications and Technology Convention, roundtable session. Indianapolis, Indiana. Retrieved from http://www.aect.org/pdf/proceedings15/2015i/15_04.pdf

Davis, A. B., Rand, R., & Seay, R. (2016). Remote proctoring: The effect of proctoring on grades. In *Advances in Accounting Education: Teaching and Curriculum Innovations* (pp. 23-50). Bingley, England: Emerald Group Publishing Limited. doi:10.1108/S1085-462220160000018002

Davis, M.R. (2014). *Adaptive Tech, Secure Browsers Aim to Curb Student Cheating*. Retrieved from http://www.edweek.org/ew/articles/2014/03/13/25cheating.h33.html?cmp=ENL-EU-NEWS2.

Dell, K. A., & Wantuch, G. A. (2017). How-to-guide for writing multiple choice questions for the pharmacy instructor. *Currents in Pharmacy Teaching and Learning*, *9*(1), 137-144. doi:10.1016/j.cptl.2016.08.036

Desai, R. L. (2015). Current issues in higher education: faculty productivity, distance education, course assessment, and online proctoring. *International Journal of Research in Business and Technology*, *6*(3), 828-833. doi:10.17722/ijrbt.v6i3.401

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*(3), 91-99.

Diedenhofen, B., & Musch, J. (2016). PageFocus: Using paradata to detect and prevent cheating on online achievement tests. *Behavior Research Methods*, 1-16. doi:10.3758/s13428-016-0800-7

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in health sciences education*, *10*(2), 133-143. doi:10.1007/s10459-004-4019-5

Dündar, S., Temel, H., & Gündüz, N. (2016). Development of a mathematical ability test: a validity and reliability study. *International Journal of Mathematical Education in Science and Technology*, *47*(7), 1061-1075. doi:10.1080/0020739X.2016.1153734

Ertürk, N. O. (2015). Testing your tests: Reliability issues of academic English exams. *International Journal of Psychology and Educational Studies*, *2*(2), 47-52. Retrieved from http://www.ijpes.com/frontend/articles/pdf/v02i02/v02i02-05.pdf

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*(7), 1095-1112. doi:10.1080/13803390500205718

Faurer, J. (2013) Grade validity of online quantitative courses. *Contemporary Issues in Education Research*, *6*(1), 93-96. doi:10.19030/cier.v6i1.7607

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*(10), 1-12. Retrieved from http://pareonline.net/getvn.asp?v=20&n=10

Fask, A., Englander, F., & Wang, Z. (2014). Do online exams facilitate cheating? An experiment designed to separate possible cheating from the effect of the online test taking environment. *Journal of Academic Ethics*, *12*(2), 101-112. doi:10.1007/s10805-014-9207-1

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191. doi:10.3758/BF03193146

Feinberg, R. A., Raymond, M. R., & Haist, S. A. (2015). Repeat testing effects on credentialing exams: Are repeaters misinformed or uninformed? *Educational Measurement: Issues and Practice*, *34*(1), 34-39. doi:10.1111/emip.12059

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. New Delhi, India: Sage Publications.

Fisher, E., McLeod, A., Savage, A., & Simkin, M. (2016). "Ghostwriters in the cloud." *Journal of Accounting Education, 34*, 59-71. doi:10.1016/j.jaccedu.2015.11.001

Foster, D., & Layman, H. (2013). Online proctoring system compared. Retrieved from https://www.researchgate.net/file.PostFileLoader.html?id...assetKey...

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (1993). *How to design and evaluate research in education*. New York, NY: McGraw-Hill.

Frankfort-Nachmias, C., & Nachmias, D. (2008). Research methods in the social sciences. New York, NY: Worth Publishers

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, *7*(1), 25-35. Retrieved from http://www.wyoaac.org/resources/Reliability%20of%20Scores%20from%20Teacher%20Made%20Tests%20-%20NCME.pdf

Gallant, T. B., Binkin, N., & Donohue, M. (2015). Students at risk for being reported for cheating. *Journal of Academic Ethics*, *13*(3), 217-228. doi:10.1007/s10805-015-9235-5

García-Cabrera, L., Ortega-Tudela, J. M., Balsas-Almagro, J. R., Ruano-Ruano, I., Peña-Hita, M. Á., & Cuevas-Martínez, J. C. (2012). New assessment tools in learning management systems. Pixel International Conference *The Future of Education*, Florence, Italy: Simonelli Editore-University Press. Retrieved from https://www.researchgate.net/profile/Lina_Garcia-Cabrera/publication/229494305_New_Assessment_Tools_in_Learning_Management_Systems/links/0912f500ee386401af000000.pdf

Glazer, S. (2013). Plagiarism and cheating. *CQ Researcher*, *23*(1), 1-28. Retrieved from http://library.cqpress.com/

Grijalva, T. C., Nowell, C., & Kerkvliet, J. (2006). Academic honesty and online courses. *College Student Journal*, *40*(1), 180. Retrieved from https://s3.amazonaws.com/academia.edu.documents/3457387/cheat_online_pap.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1511759085&Signature=bZ5alQPgmTLr6j7Y%2F1yYaIS8JvY%3D&response-content-disposition=inline%3B%20filename%3DAcademic_Honesty_and_Online_Courses.pdf

Gustafson, C. R. (2002). Online assessment-faculty and student perspectives. Retrieved from http://wwwic.cs.ndsu.nodak.edu/conferences/beyond2002/files

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, *15*(3), 309-333. doi:10.1207/S15324818AME1503_5

Hanson, P., & Robson, R. (2004). Evaluating course management technology: A pilot study. *EDUCAUSE Center for Applied Research, Research Bulletin*, *24*. Boulder, CO: EDUCAUSE. Retrieved from https://www.educause.edu/ir/library/pdf/ERB0424.pdf

Hameed, I. A. (2016). A fuzzy system to automatically evaluate and improve fairness of multiple-choice questions (MCQs) based exams. In *Proceedings of the 8th International Conference on Computer Supported Education* (pp. 476-481). doi:10.5220/0005897204760481

Harding, T. S. (2001). Useful approaches to preventing academic dishonesty in the classroom. Retrieved from http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1066&context=mate_fac

Harmon, O. R., & Lambrinos, J. (2008). Are online exams an invitation to cheat? *The Journal of Economic Education*, *39*(2), 116-125. doi:10.3200/JECE.39.2.116-125

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: a meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373. doi:10.1037/0021-9010.92.2.373

Hayes, D., Hurtt, K., & Bee, S. (2006). The war on fraud: Reducing cheating in the classroom. *Journal of College Teaching & Learning (TLC)*, *3*(2). doi:10.19030/tlc.v3i2.1742

Higher Education Opportunity Act (HEOA). (2008). Public Law No.110-315,112 Stat. §§3078-3508.

Hodgkinson, T., Curtis, H., MacAlister, D., & Farrell, G. (2016). Student academic dishonesty: The potential for situational prevention. *Journal of Criminal Justice Education*, *27*(1), 1-18. doi:10.1080/10511253.2015.1064982

Hylton, K., Levy, Y., & Dringus, L. P. (2016). Utilizing webcam-based proctoring to deter misconduct in online exams. *Computers & Education*, *92*, 53-63. doi:10.1016/j.compedu.2015.10.002

Imran, A. M., & Nordin, M. S. (2013). Predicting the underlying factors of academic dishonesty among undergraduates in public universities: A path analysis approach. *Journal of Academic Ethics*, *11*(2), 103-120. doi:10.1007/s10805-013-9183-x

Instructional Technology Council. (2016). *ITC Annual National eLearning Report.2015 Survey Results.* Retrieved from http://www.itcnetwork.org/resources/1439-itc-2015-distance-learning-survey-results.html

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

Itmazi, J. A., Megías, M. G., Paderewski, P., & Vela, F. L. G. (2005). A comparison and evaluation of open source learning management systems. Paper presented at IADIS International Conference of Applied Computing, Algarve, Portugal. Retrieved from https://pdfs.semanticscholar.org/32c3/63a0ac4b200e99f988739f4b3d31b25b529c.pdf

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, *118*(3), 843-877. doi:10.1162/00335530360698441

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test… or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, *26*(2), 307-329. doi:10.1007/s10648-013-9248-9

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology, 33*(4), 410-422. doi:10.1080/0144929X.2012.710647

Jones, C. (2010). Archival data: Advantages and disadvantages for research in psychology. *Social and Personality Psychology Compass*, *4*(11), 1008-1017. doi:10.1111/j.1751-9004.2010.00317.x

Jones, I. S., Blankenship, D., & Hollier, G. (2013). Am I cheating? An analysis of online students' perceptions of their behaviors and attitudes. *Psychology Research, 3*(5), 261-271. Retrieved from http://asbbs.org/files/ASBBS2013V1/PDF/J/Jones_Blankenship_Hollier

Kainz, O., Cymbalák, D., & Jakab, F. (2015). Adaptive web-based system for examination with cheating prevention mechanism. *Lecture Notes on Software Engineering*, *3*(2), 90-94. Retrieved from http://www.lnse.org/vol3/172-P003.pdf

Kaya, Z., Tan, S. (2014). New trends of measurement and assessment in distance education. *Turkish Online Journal of Distance Education, 15*(1), 206-217. doi:10.17718/tojde.30398

Khan, Z. R., & Balasubramanian, S. (2012). Students go click, flick and cheat... e-cheating, technologies and more. *Journal of Academic and Business Ethics*, *6*, 1-26. Retrieved from https://pdfs.semanticscholar.org/8f9f/1116ac277f8a4390f09a78c3dca675bb9f1d

Kibble, J. D. (2017). Best practices in summative assessment. *Advances in Physiology Education*, *4*(1), 110-119. doi:10.1152/advan.00116.2016

Kim, Y., & Steiner, P. (2016). Quasi-experimental designs for causal inference. *Educational Psychologist*, *51*(3), 395-405. doi:10.1080/00461520.2016.1207177

King, C. G., Guyette Jr, R. W., & Piotrowski, C. (2009). Online exams and cheating: An empirical analysis of business students' views. *Journal of Educators Online*, *6*(1), 1-11. doi:10.9743/JEO.2009.1.5

Koschel, T. L., Young, J. C., & Navalta, J. W. (2017). Examining the impact of a university-driven exercise programming event on end-of-semester stress in students. *International Journal of Exercise Science*, *10*(5), 754. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5609663/

Krathwohl, D. R. (2002). A revision of Bloom’s taxonomy: An overview. *Theory into practice*, *41*(4), 212-218. doi:10.1207/s15430421tip4104_2

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*(7), 883-897. doi:10.1080/02602938.2014.956683

Lang, J. (2013). *Cheating Lessons: Learning from Academic Dishonesty*. Cambridge, MA: Harvard University Press.

Lee, K. S., Osborne, R. E., & Carpenter, D. N. (2010). Testing accommodations for university students with AD/HD: Computerized vs. paper-pencil/regular vs. extended time. *Journal of Educational Computing Research*, *42*(4), 443-458. doi:10.2190/EC.42.4.e

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

Lewellyn, P., & Rodriguez, L. (2015). Does academic dishonesty relate to Fraud Theory? A comparative analysis. *American International Journal of Contemporary Research, 5*(3), 1-6. Retrieved from http://www.aijcrnet.com/journals/Vol_5_No_3_June_2015/1.pdf

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215-233. doi:10.1002/tea.21299

London, M. (2014). The History and Status of Remote Proctoring. *Distance Learning*, *11*(1), 61–62. Retrieved from https://eds-a-ebscohost-com.ezp.waldenulibrary.org/eds/pdfviewer/pdfviewer?vid=4&sid=9d49d213-a651-477f-b526-e065d0b9a343%40sessionmgr4008

Lonn, S., & Teasley, S. D. (2009). Saving time or innovating practice: Investigating perceptions and uses of Learning Management Systems. *Computers & Education*, *53*(3), 686-694. doi:10.1016/j.compedu.2009.04.008

Lushene, R. E., O'Neil Jr, H. F., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment*, *38*(4), 353-361. doi:10.1016/j.compedu.2009.04.008

MacGregor, J., & Stuebs, M. (2012). To cheat or not to cheat: Rationalizing academic impropriety. *Accounting Education*, *21*(3), 265-287. doi:10.1080/09639284.2011.617174

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*(4), 117. Retrieved from https://eds-a-ebscohost-com.ezp.waldenulibrary.org/eds/pdfviewer/pdfviewer?vid=7&sid=9d49d213-a651-477f-b526-e065d0b9a343%40sessionmgr4008

Makar, K., & Ben-Zvi, D. (2011). The role of context in developing reasoning about informal statistical inference. *Mathematical Thinking and Learning*, *13*(1-2), 1-4. doi:10.1080/10986065.2011.538291

Malesky, L. A., Baley, J., & Crow, R. (2016). Academic Dishonesty: Assessing the Threat of Cheating Companies to Online Education. *College Teaching*, *64*(4), 178-183. doi:10.1080/10986065.2011.538291

Malgwi, C. A., Rakovski, C. C. (2008). Behavioral implications and evaluation of academic fraud risk factors. *Journal of Forensic Accounting*, *1*(2), 1-37. Retrieved from https://www.researchgate.net/profile/Carter_Rakovski/publication

Mayer, A., & Krampen, G. (2015). Equivalence of computerized versus paper-and-pencil testing of information literacy under controlled versus uncontrolled conditions: An experimental study. Presentation at 13th European Conference on Psychological Assessment, July, Zurich, Switzerland. Retrieved from https://www.zpid.de/pub/research/2015_Mayer-Krampen_Equivalence.pdf

Mazzeo, J., & Harvey, A. L. (1988). The equivalence of scores from automated and conventional educational and psychological tests. *ETS Research Report Series*, *1988*(1), 1-27. doi:10.1002/j.2330-8516.1988.tb00277.x

McCabe, D., Butterfield, L., Trevino, L. (2012). *Cheating in College: Why Students do it and What Educators can do about it.* Baltimore, MD: The John Hopkins University Press.

McCaslin, S., & Brown, F. (2015). Case study: Challenges and issues in teaching fully online mechanical engineering courses. In *New Trends in Networking, Computing, E-learning, Systems Sciences, and Engineering* (pp. 575–579). Cham, Switzerland: Springer. Retrieved from https://s3.amazonaws.com/academia.edu.documents/46425908/Case_Study_Challenges_and_Issues_in_Teac20160612-12327-aphyfe.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1511761352&Signature=vu9Wow5ciBI6ergeYG05NwhUO70%3D&response-content-disposition=inline%3B%20filename%3DCase_Study_Challenges_and_Issues_in_Teac.pdf

McGee, P. (2013). Supporting Academic Honesty in Online Courses. *Journal of Educators Online*, *10*(1), 1-32. doi:10.9743/JEO.2013.1.6

McLeod, I., Zhang, Y., & Yu, H. (2003). Multiple-choice randomization. *Journal of Statistics Education*, *11*(1), 8-9. Retrieved from cww2.amstat.org/publications/jse/v11n1/mcleod.html

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*(3), 449-458. Retrieved from https://www.researchgate.net/profile/Alan_Mead/publication/232485286

Mertler, C. A., & Vannatta, R. A. (2002). *Advanced and multivariate statistical methods.* Los Angeles, CA: Pyrczak.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin*, *89*(2), 191-200. doi:10.1037/0033-2909.89.2.191

Miller, R., & Hollist, C. (2007). Attrition bias. *Encyclopedia of Measurement and Statistics*, *3*(1), 57-60. Retrieved from http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1044&context=famconfacpub

Milone, A. S., Cortese, A. M., Balestrieri, R. L., & Pittenger, A. L. (2017). The impact of proctored online exams on the educational experience. *Currents in Pharmacy Teaching and Learning*, *9*(1), 108-114. doi:10.1016/j.cptl.2016.08.037

Moodle. (2016a, May 10). *Quiz activities*. Retrieved from https://docs.moodle.org/30/en/Quiz_activity

Moodle. (2016b, September 8). *Question types*. Retrieved from https://moodle.org/plugins/browse.php?list=category&id=29

Moodle. (2016c, April 4). *Moodle Mobile*. Retrieved from https://docs.moodle.org/dev/Accessibility

Moodle. (2016d, August 23). *Accessibility*. Retrieved from https://docs.moodle.org/dev/Accessibility

Moodle. (2016e, April 17). *Learning Analytics*. Retrieved from https://docs.moodle.org/30/en/Learning_analytics

Moodle. (2015, January15). *Effective quiz practices: quiz security and cheating*. Retrieved from https://docs.moodle.org/25/en/Effective_quiz_practices

Moodle. (2012, June 21). *History of the Moodle quiz and question bank*. Retrieved from https://docs.moodle.org/dev/History_of_the_Moodle_quiz_and_question_bank

Moore, P., Head, J., & Griffin, R. (2017). Impeding students' efforts to cheat in online classes. *Journal of Learning in Higher Education*, *13*(1). Retrieved from https://files.eric.ed.gov/fulltext/EJ1139692.pdf

Moten, J., Fitterer, A., Brazier, E., Leonard, J., & Brown, A. (2013). Examining online college cyber cheating methods and prevention measures. *Electronic Journal of e-Learning, 11*(2), 139-146. Retrieved from https://files.eric.ed.gov/fulltext/EJ1012879.pdf

Murphy, K., Myors, B, & Wolach, A. (2014). *Statistical Power Analysis. A Simple and General Model for Traditional and Modern Hypothesis Tests.* New York, NY: Routledge Taylor & Francis Group.

Nash, J. A. (2015). Future of online education in crisis: A call to action. *Turkish Online Journal of Educational Technology, 14*(2), 80-91. Retrieved from https://files.eric.ed.gov/fulltext/EJ1057370.pdf

Nash, J. M., & Krauss, K. E. M. A. (2015, July). *Method for aligning MCQ assessment with cognitive skills and learning objectives.* Paper presented on 44th Conference of the Southern African computer lecturers' association, Johannesburg, South Africa. Retrieved from https://www.researchgate.net/publication/281060985

Nilsson, L. E. (2016). Technology as a Double-Edged Sword: A Promise Yet to Be Fulfilled or a Vehicle for Cheating? In *Handbook of academic integrity* (pp. 607-620). Singapore, Singapore: Springer. doi: 0.1007/978-981-287-098-8_21

New Media Consortium. (2015). *The NMC Horizon Report: 2015 Higher Education Edition.* Retrieved from http://cdn.nmc.org/media/2015-nmc-horizon-report-HE-EN.pdf

New Media Consortium. (2016). *The NMC Horizon Report: 2016 Higher Education Edition.* Retrieved from http://cdn.nmc.org/media/2016-nmc-horizon-report-he-EN.pdf

New Media Consortium. (2017). *The NMC Horizon Report: 2017 Higher Education Edition.* Retrieved from http://cdn.nmc.org/media/2017-nmc-horizon-report-he-EN.pdf

Nkundabanyanga, S., Omagor, C., & Nalukenge, I. (2014). Correlates of academic misconduct and CSR proclivity of students. *Journal of Applied Research in Higher Education*, *6*(1), 128-148. doi:10.1108/JARHE-05-2012-0016

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. Retrieved from https://arxiv.org/ftp/arxiv/papers/1508/1508.05699.pdf

Odegard, T. N., & Koen, J. D. (2007). “None of the above” as a correct and incorrect alternative on a multiple-choice test: implications for the testing effect. *Memory*, *15*(8), 873-885. Retrieved from https://www.researchgate.net/profile/Joshua_Koen/publication/5880101

Olivero, J. (2013) Frequency of student cheating on online test examinations. *National Social Science Technology Journal*, *3*(2). Retrieved from http://www.nssa.us/tech_journal/volume_3-2/vol3-2_article4.htm

O'Reilly, G. & Creagh, J. (2016). A categorization of online proctoring. In *Proceedings of Global Learn-Global Conference on Learning and Technology* (pp. 542-552). Limerick, Ireland: Association for the Advancement of Computing in Education (AACE). Retrieved from https://www.learntechlib.org/p/172801/

Patelis, T. (2000). An Overview of Computer-Based Testing. *College Entrance Examination Board Research Notes*, *9*, 1-6. Retrieved from http://files.eric.ed.gov/fulltext/ED562592.pdf

Paullet, K., Douglas, D., & Chawdhry, A. (2015). Student perspectives of cheating in online classes. *Issues in Information Systems*, *16*(4), 215-223. Retrieved from http://www.iacis.org/iis/2015/4_iis_2015_215-223.pdf

Pittman, V. (2015). First System, Best System: The Proctored Examination. *New Horizons in Adult Education and Human Resource Development*, *27*(1), 44-50. doi:10.1002/nha3.20093

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, *31*(4), 463-479. Retrieved from https://www.researchgate.net/profile/Steffi_Pohl/publication/249795570

Prince, D. J., Fulton, R. A., & Garsombke, T. W. (2009). Comparisons of proctored versus non-proctored testing strategies in graduate distance education curriculum. *Journal of College Teaching and Learning*, *6*(7), 51-61. Retrieved from https://www.learntechlib.org/p/108148/

Randall, J. G., & Villado, A. J. (2016). Take two: Sources and deterrents of score change in employment retesting. *Human Resource Management Review*, *27*(3), 536-553. doi:10.1016/j.hrmr.2016.10.002

Ravasco, G. G. (2012). Technology-aided cheating in open and distance e-learning. *Asian Journal of Distance Education*, *10*(2), 71-77. Retrieved from http://www.asianjde.org/2012v10.2.Ravasco.pdf

Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology*, *60*(2), 367-396. doi:10.1111/j.1744-6570.2007.00077.x

Raymond, C., & Usherwood, S. (2013). Assessment in simulations. *Journal of Political Science Education*, *9*(2), 157-167. doi:10.1080/15512169.2013.770984

Redecker, C., & Johannessen, Ø. (2013). Changing assessment—Towards a new assessment paradigm using ICT. *European Journal of Education*, *48*(1), 79-96. Retrieved from https://www.mycota.ca/assets/uploads/documents/eAssessment.pdf

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3-13. Retrieved from http://www.highpoint.edu/citl/files/2017/06/Three_Options_Are_Optimal_for_MCQ_Rodriguez_2005.pdf

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210. doi:10.1111/j.1745-6916.2006.00012.x

Rogers, C. F. (2006). Faculty perceptions about e-cheating during online testing. *Journal of Computing Sciences in Colleges*, *22*(2), 206-212. Retrieved from https://www.researchgate.net/profile/Camille_Rogers2/publication/262311152_Faculty_perceptions_about_e-cheating_during_online_testing/links/5751b50f08ae02ac127786b8.pdf

Romero, C., Ventura, S., & De Bra, P. (2009). Using mobile and web-based computerized tests to evaluate university students. *Computer Applications in Engineering Education*, *17*(4), 435-447. Retrieved from https://pdfs.semanticscholar.org/90f5/9dd323ef9e241b898a4b9d513e5fe4f88b3c.pdf

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing" intermediate constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment*, *4*(6), 1-48. Retrieved from https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1653/0

Schlegel, R., & Gilliland, K. (2007). Development and quality assurance of computer-based assessment batteries. *Archives of Clinical Neuropsychology*, *22*, 49-61. doi:10.1016/j.acn.2006.10.005

Schumaker, R., & Lomax, R (2004). *A Beginner's Guide to Structural Equation Modeling*. Mahwah, NJ: Erlbaum.

Seidman, A. (2005). *College Student Retention: Formula for Student Success*. Westport, CT: Greenwood Publishing Group.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Wadsworth Cengage learning.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*(484), 1334-1344. doi:10.1198/016214508000000733

Shadish, W. R. (2011). Randomized controlled studies and alternative designs in outcome studies: Challenges and opportunities. *Research on Social Work Practice*, *21*(6), 636-643. doi:10.1177/1049731511403324

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, *78*(3), 481-497. doi:10.1007/s11336-012-9311-3

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), 1-19. doi:10.1111/jcal.12172

Simpson, E., & Yu, K. (2012). Closer to the Truth: Electronic Records of Academic Dishonesty in an Actual Classroom Setting. *Ethics & Behavior*, *22*(5), 400-408. doi:10.1080/10508422.2012.702514

Siniver, E. (2013). Cheating on exams: The case of Israeli students. *College Student Journal*, *47*(4), 593-604. Retrieved from http://www.ingentaconnect.com/content/prin/csj/2013/00000047/00000004

Sivula, M., & Robson, E. (2015). E-testing in graduate courses: Reflective practice case studies. *MBA Faculty Conference Papers & Journal Articles*, *87*, 1-8. Retrieved from http://scholarsarchive.jwu.edu/cgi/viewcontent.cgi?article=1086&context

Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational researcher*, *37*(1), 5-14. doi:10.3102/0013189X08314117

Slepkov, A. D., & Shiell, R. C. (2014). Comparison of integrated testlet and constructed-response question formats. *Physical Review Special Topics-Physics Education Research*, *10*(2), 020120. doi:10.1103/PhysRevSTPER.10.020120

Smatter Balanced Assessment Consortium. (2016). *2013-2014 Technical report.* Retrieved from http://www.smarterbalanced.org/wp-content/uploads/2015/08/2013-14_Technical_Report.pdf

Smatter Balanced Assessment Consortium. (2014). Field Test: Automated Scoring Research Studies (Smarter Balanced RFP 17). Retrieved from http://www.smarterapp.org/documents/FieldTest_AutomatedScoringResearchStudies.pdf

Smarter Balanced Assessment Consortium. (n.d.). *Higher Ed Approved.* Retrieved from http://www.smarterbalanced.org/about/higher-education/ on August 30, 2016.

Song, M. K., & Ward, S. E. (2015). Assessment effects in educational and psychosocial intervention trials: An important but Often-overlooked problem. *Research in Nursing & Health, 38*(3), 241-247. doi:10.1002/nur.21651/full

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*(9), 641-656. doi:10.1037/h0063404

Srikanth, M., & Asmatulu, R. (2014). Modern cheating techniques, their adverse effects on engineering education and preventions. *International Journal of Mechanical Engineering Education, 42*(2), 129-140. doi:10.7227/IJMEE.0005

Staats, S., & Hupp, J. M. (2012). An examination of academic misconduct intentions and the ineffectiveness of syllabus statements. *Ethics & Behavior*, *22*(4), 239-247. doi:10.1080/10508422.2012.661313

Stack, S. (2015). The Impact of exam environments on student test scores in online courses. *Journal of Criminal Justice Education*, *26*(3), 1–10. doi:10.1080/10511253.2015.1012173

Stapleton, C. D. (1997). Basic concepts in exploratory factor analysis (EFA) as a tool to evaluate score validity: A right-brained approach. Retrieved from http://files.eric.ed.gov/fulltext/ED407416.pdf

Stuber-McEwen, D., Wiseley, P., & Hoggatt, S. (2009). Point, click, and cheat: Frequency and type of academic dishonesty in the virtual classroom. *Online Journal of Distance Learning Administration*, *12*(3), 1-10. Retrieved from http://www.westga.edu/~distance/ojdla/fall123/stuber123.html

Sukamolson, S. (2010). Fundamentals of quantitative research. *Language Institute, Chulalongkorn University*. Retrieved from

http://www.culi.chula.ac.th/Research/e-Journal/bod/Suphat%20Sukamolson.pdf

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics*. Boston, MA: Pearson.

Tal, I. R., Akers, K. G., & Hodge, G. K. (2008). Effect of paper color and question order on exam performance. *Teaching of Psychology*, *35*(1), 26-28. doi:10.1080/00986280701818482

Tavakol, S., Dennick, R., & Tavakol, M. (2011). Psychometric properties and confirmatory factor analysis of the Jefferson Scale of Physician Empathy. *BMC medical education*, *11*(1), 54-67. doi:10.1186/1472-6920-11-54

Templin, J. (2015). Item Response Theory. *The Encyclopedia of Adulthood and Aging*, 1-5. doi:10.1002/9781118521373.wbeaa320

Templer, K. J., & Lange, S. R. (2008). Internet testing: Equivalence between proctored lab and unproctored field conditions. *Computers in Human Behavior*, *24*(3), 1216-1228. doi:10.1016/j.chb.2007.04.006

Thompson, C. B., & Panacek, E. A. (2006). Research study designs: experimental and quasi-experimental. *Air medical journal*, *25*(6), 242-246. doi:10.1016/j.amj.2006.09.001

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2015). Rethinking assessment in a digital age: opportunities, challenges and risks. *British Educational Research Journal*, *42*(3), 454-476. doi:10.1002/berj.3215

The Institute for College Access and Success. (April, 2016). On the verge: Costs and tradeoffs facing community college students. Retrieved from http://ticas.org/sites/default/files/pub_files/on_the_verge.pdf

Tindell, D. R., & Bohlander, R. W. (2012). The use and abuse of cell phones and text messaging in the classroom: A survey of college students. *College Teaching*, *60*(1), 1-9. doi:10.1080/87567555.2011.604802

Tinkelman, D. (2012). Using auditing concepts to discourage college student academic misconduct and encourage engagement. *Journal of Academic and Business Ethics*, *5*, 1-28. Retrieved from http://www.aabri.com/manuscripts/11958.pdf

Tinto, V. (2012). *Completing College: Rethinking Institutional Action*. Chicago, IL: University of Chicago Press.

Towns, M. H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education*, *91*(9), 1426-1431. doi:10.1021/ed500076x

Tractenberg, R. E., Gushta, M. M., Mulroney, S. E., & Weissinger, P. A. (2013). Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in Health Sciences Education*, *18*(5), 945-961. doi:10.1007/s10459-012-9434-4

Trenholm, S. (2007). A review of cheating in fully asynchronous online courses: A math or fact-based course perspective. *Journal of Educational Technology Systems*, *35*(3), 281-300. doi:10.2190/Y78L-H21X-241N-7Q02

Trochim, W. M. (1986). *Advances in Quasi-experimental Design and Analysis*. San Francisco, CA: Jossey-Bass Inc Pub.

Trochim, W. M., & Donnelly, J. P. (2006). *Research Methods: Knowledge Base*. Retrieved from https://www.socialresearchmethods.net/kb/consthre.php

Trochim, W.M., Donnelly, J.P., Arora, K. (2016). *Research Methods: The Essential Knowledge Base.* Retrieved from https://cengagebrain.vitalsource.com/#/books/9781305445185/cfi/51!/4/4

Tuckman, B. W. (2003). The effect of learning and motivation strategies training on college students' achievement. *Journal of College Student Development*, *44*(3), 430-437. doi:10.1353/csd.2003.0034

Van Der Vleuten, C. P. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education*, *1*(1), 41-67. doi:10.1007/BF00596229

Varble, D. (2014). Reducing cheating opportunities in online tests. *Atlantic Marketing Journal*, *3*(2), 131-149. Retrieved from http://digitalcommons.kennesaw.edu/cgi/viewcontent.cgi?article=1110&context=amj

Villado, A. J., Randall, J. G., & Zimmer, C. U. (2016). The effect of method characteristics on retest score gains and criterion-related validity. *Journal of Business and Psychology*, *31*(2), 233-248. doi:10.1007/s10869-015-9408-7

Wachenheim, C. J. (2009). Final exam scores in introductory economics courses: Effect of course delivery method and proctoring. *Review of Agricultural Economics*, *31*(3), 640-652. doi:10.1111/j.1467-9353.2009.01458.x

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*(1), 32-42. doi:10.1108/JARHE-02-2014-0024

Weatherly, M., Jennings, S., & Wilson, S. (2015). Online Integrity: Student Authentication in an Online Course. *Journal of Research in Business Information Systems*, 57-72. Retrieved from http://scholarworks.sfasu.edu/cgi/viewcontent.cgi?article=1432&context=forestry

Western Cooperative for Educational Telecommunication. (2009, June). *Best practice strategies to promote academic integrity in online education*. Retrieved from http://wcet.wiche.edu/sites/default/files/docs/resources/Best-Practices-Promote-Academic-Integrity-2009.pdf

Widianingsih, L. P. (2013). Students Cheating Behaviors: The Influence of Fraud Triangle. *Business Economic Res*, *2*(2), 252-261. Retrieved from http://www.sibresearch.org/uploads/2/7/9/9/2799227/riber_b13-134_252-260.pdf

Wildgrube, W. (1982). *Computerized Testing in the German Federal Armed Forces: Empirical Approaches*. ERIC Clearinghouse.

Williams Jr, T. O., Fall, A. M., Eaves, R. C., Darch, C., & Woods-Groves, S. (2007). Factor Analysis of the KeyMath—Revised Normative Update Form A. *Assessment for Effective Intervention*, *32*(2), 113-120. doi:10.1177/15345084070320020201

Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting Test Tampering Using Item Response Theory. *Educational and Psychological Measurement*, doi:10.0013164414568716.

Xu, X., Kauer, S., & Tupy, S. (2016). Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Scholarship of Teaching and Learning in Psychology*, *2*(2), 147. doi:10.1037%2Fstl0000062

Zenisky, A. L., & Sireci, S. G. (2007). A Summary of the research on the effects of test accommodations: 2005-2006. Technical Report 47. *National Center on Educational Outcomes, University of Minnesota*. Retrieved from http://files.eric.ed.gov/fulltext/ED499407.pdf

Zito, N., & McQuillan, P. J. (2010). Cheating themselves out of an education: assignments that promote higher-order thinking and honesty in the middle grades. *Middle School Journal*, *42*(2), 6-16. doi:10.1080/00940771.2010.11461752

Appendix A: The Study’s Opportunity Reduction Mechanisms

Reduction Mechanism	Purpose	Influence on other fraud triangle factors	Neutralization of negative side-effects
1. Synchronous testing	(a) Eliminates the possibility of taking the same exam at different time (Stack, 2015) (b) Prevents circulation of exam items (Stack, 2015) (c) Reduces a need for numerous different exam’s versions.	(a) May trigger *rationalization* “I did not know that the quiz would not be open during the whole day and did not allocate enough time to study” (b) May increase *opportunity* for collusion when two or more students work on the same exam side-by-side	(a) Early exams’ dates/times announcements (Malgwi & Rakovski, 2008; Tinkelman, 2012); frequent reminders about the tests’ days (b) Randomization of questions, one question per page, blocked backtracking (Beck, 2014; Stack, 2015)
2. Limited test taking time	(a) Does not allow time for looking up the answers on the Internet, in the printed sources, by texting/emailing friends (Beck, 2014; Stack, 2015; Varble, 2014) (b) Does not allow time for collusions (Stack, 2015)	(a) May increase *rationalization* “the test is too hard” (b) May trigger a need to cheat due to fear of getting bad grades	(a) Careful identification of the period sufficient to complete exams without rushing (Hodgkinson, 2016) (b) Practice tests with the same number of minutes per question
3. Blocked backtracking 4. One question per page	(a) Prevents collaboration and collusion (Beck, 2014; Stack, 2015) (b) With restricted time, may prevent dissemination of questions (Beck, 2014)	May increase *rationalization “*the test is not fair for online students”	Uniform curriculum and assessment across all delivery modes (Malgwi & Rakovski, 2008) Clear course expectations (Tinkelman, 2012), discussions about the importance of credibility of exams, high standards, and effective learning and instruction. Practice tests with blocked backtracking
5. High order thinking exam questions	Makes looking up the answers to exam questions on the internet/books pointless (Ladyshewsky, 2015; Varble, 2014)	(a) May increase *rationalization* “the test is too hard” (b) May trigger a need to cheat due to fear of getting bad grades	(a) High order thinking questions included in all course assignments: homework, discussions, quizzes, study guides, practice tests, exams (b) Discussions about importance of high order cognitive skills

(table continues)

Reduction Mechanism	Purpose	Influence on other fraud triangle factors	Neutralization of negative side-effects
6. Randomization of questions and responses	Prevents cheating when students sit next to each other (Beck, 2014; Stack, 2015)	May increase *rationalization* “the test is too hard”	Creation of questions in accordance with the best practices (Harding, 2001; Tinkelman, 2012; Xu et al., 2016)
7. Deferred feedback	Prevents distribution of correct answers, hints and solutions (Beck, 2014)	May increase *rationalization* “the test is not fair”	Individual feedback on students’ answers
8. Multiple versions of the same exam	Prevents cheating due dissemination of test items (Tinkelman, 2016)	May trigger *rationalization* “the test is not fair” if different versions of tests are not equivalent	Equivalency of all versions of the exam
9. Inaccessibility of the exams after they are submitted	Prevents distribution of tests items (Beck 2014; Stack, 2015)	May trigger *rationalization* “it is not fair” when students cannot see their tests	Clear explanations why the exams are not accessible Individual feedback on students’ answers

Appendix B: The Study's Rationalization Reduction Mechanisms

Reduction mechanisms	Purpose	Influence on other fraud triangle factors	Neutralization of negative side-effects
1. Clear description of what constitute academic misconduct in the class syllabus (Beck, 2014; Tinkelman, 2012)	Prevents rationalization "I did not know that it was cheating" (Tinkelman, 2012)	May reduce perceived opportunities to cheat (Tinkelman, 2012)	N/A
2. Clearly stated course expectations (Tinkelman, 2012)	Reduces rationalization "The course is too hard" (Malgwi & Rakovski, 2008; Tinkelman, 2012)	May reduce "fear of undesired grades" *need* (Malgwi & Rakovski, 2008; Tinkelman, 2012)	N/A
3. Clearly stated severe consequences of cheating (Siniver, 2013; Tinkelman, 2012)	Reduces rationalization "instructors do not care about cheating and do not prevent it" (Tinkelman, 2012)	May reduce perceived opportunities to cheat (Ladyshewsky, 2015; Siniver, 2013)	N/A
4. Faculty intolerance to cheating (Tinkelman, 2012)			
5. Cheating warning statement (Beck, 2014)	Prevents rationalization "Instructors do not care about cheating and do not try to prevent it" (Tinkelman, 2012)	May reduce perceived opportunities to cheat (Beck, 2014; Corrigan-Gibbs et al., 2015)	N/A
6. Timely feedback (Tinkelman, 2012)	Reduces rationalization "The course is too hard" and "instructors do not care" (Tinkelman, 2012)	(a) May reduce "fear of undesired grades" *need* (Tinkelman, 2012)	
		(b) May increase opportunity to cheat using instructor's feedback	High order thinking questions (Ladyshewsky, 2015) and inaccessibility of online feedback (Stack, 2015)

(table continues)

Reduction mechanisms	Purpose	Influence on other fraud triangle factors	Neutralization of negative side-effects
7. Fairness of the tests (Harding, 2001) 8. Fair grading based on automatic scoring/ well-designed grading rubrics (Harding, 2001)	(a) Reduces rationalization “The test is too hard” (Tinkelman, 2012). (b) Prevents rationalization “The instructor is a hard grader” (Tinkelman, 2012)	May reduce “fear of undesired grades” *need* (Tinkelman, 2012)	N/A
9. The list of the concepts covered on the test posted in advance (Tinkelman, 2012)	(a) Reduces rationalization “The test is too hard” (Tinkelman, 2012)	(a) May reduce “fear of undesired grades” *need* (Tinkelman, 2012) (b) My increase *opportunity* the use of cheat sheets	Avoidance of using exam questions focused on remembrance
10. Availability of old tests with solutions (Tinkelman, 2012)	Reduces rationalization “The test is too hard” (Tinkelman, 2012)	(a) May reduce “fear of undesired grades” *need* (Tinkelman, 2012) (b) May increase cheating *opportunity* to use the old tests’ solutions	New versions of exams each semester Unique high order thinking test items Restricted test time
11. Online practice tests 12. Video recordings of exams’ study sessions	(a) Reduces rationalization “The test is too hard” (Tinkelman, 2012) (b) Prevents rationalization “The instructor is a hard grader” (Tinkelman, 2012)	(a) May reduce “fear of undesired grades” *need* (Tinkelman, 2012). (b) May increase cheating opportunity to use online practice tests and study sessions’ videos during exams	Inaccessibility of online practice tests/videos during exams
13. Reference sheet with formulas allowed during exams (Tinkelman, 2012)	Reduces rationalization “The course is too hard” (Tinkelman, 2012)	(a) May reduce “fear of undesired grades” need (b) Prevents cheating opportunity to look up the formulas on the internet	N/A

(table continues)

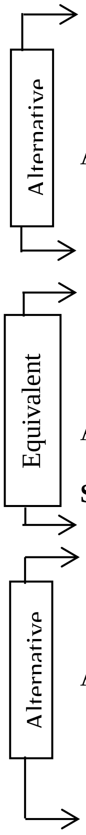
Reduction mechanisms	Purpose	Influence on other fraud triangle factors	Neutralization of negative side-effects
14. Well-developed curriculum with real-world applications relevant to students' interests (Harding, 2001; Tinkelman, 2012)	Reduces rationalization "I do not need the material covered in the course" (Tinkelman, 2012)	May reduce "fear of undesired grades" need	N/A
15. Creating a classroom atmosphere of mutual respect and trust (Lang, 2013; MacGregor & Stuebs, 2012; Tinkelman, 2012)	May reduce all rationalizations (MacGregor & Stuebs, 2012; Tinkelman, 2012)	May reduce perceived opportunities and needs to cheat (Malgwi & Rakovski, 2008; Tinkelman, 2012)	N/A
16. Emphasis on the importance of acquired knowledge (Tinkelman, 2012)	Reduces rationalization "I do not need the material covered in the course" (MacGregor & Stuebs, 2012)	May reduce perceived opportunities to cheat	N/A
17. Pedagogical uniformity and consistency in content delivery, administering the exams, and grading in all sections of the course (Malgwi & Rakovski, 2008)	Reduces rationalizations "the course/test is too hard," "the instructor is a hard grader" (Malgwi & Rakovski, 2008)	May increase cheating opportunity "sharing information about exam content with friends in other course sections"	Synchronous testing (Stack, 2015)
18. Emphasis on the importance of ethical behavior	Reduces rationalization "everyone does it" (Tinkelman, 2012)		N/A

Note. N/A= Negative effects were not identified.

Appendix C: Samples of Short-Answer Questions

The list of all needed formulas was provided on each exam.

Assigned points: 3
Cognitive process: 2.Generalities' Remembrance, 3.Comprehension, 4. Analysis, 5. Organizing, 6. Application (Darwazeh & Branch, 2015).
Knowledge: A. Factual, B. Conceptual, C. Principles, D. Procedural
Alignment Code: B2, B3, C4, D5, D6



Set 1 (Learning Objective: Calculate a CI for the population proportion)

Proctored Exam: Compute a 95% confidence interval for the population proportion of premature births based on a random sample of 181 mothers with the sample proportion of 0.09. Input your answer in the blank provided below as a three-part inequality with the correct symbol between the endpoints.

Answer: ______________________

Unproctored Exam: Calculate a 95% confidence interval for the population proportion of successful students (GPA >3.0) based on a random sample of 625 students with the sample proportion of 0.18. Input your answer in the blank provided below as a three-part inequality with the correct symbol between the endpoints.

Answer: ______________________

Set 2 (Learning Objective: Calculate a CI for the population mean)

Unproctored Exam: Calculate a 95% confidence interval for the population mean of senior swimmers' age based on a random sample of 52 swimmers with the sample mean of 68.04 and sample standard deviation of 10.31. Input your answer as a three-part inequality with the correct symbol between the endpoints in the blank provided below.

Answer: ______________________

Proctored Exam: Calculate a 95% confidence interval for the population mean of figure skaters' score based on a random sample of 115 figure skaters with the sample mean of 72.22 and sample standard deviation of 13.92. Input your answer as a three-part inequality with the correct symbol between the endpoints in the blank provided below.

Answer: ______________________

Assigned points: 1
Cognitive process: 2. Generalities' Remembrance, 3.Comprehension (Darwazeh & Branch, 2015).
Knowledge: A. Factual, B. Conceptual, D. Procedural
Alignment code: A2, B3, D3

A sample of a short-answer question that is the same on all four exams

(Learning Objective: Calculate the Margin of Error of the CI)

Calculate the margin of error of the CI [65.168, 70.908]. Insert your answer with the correct notation in the blank provided below.

Answer: ______________________

Appendix D: Samples of Drop-Down Questions

Assigned points: 4
Cognitive process: 1.Comprehension, 2.Analysis, 3. Organizing, 4. Application, 5. Evaluation (Darwazeh & Branch, 2015).
Knowledge dimension: B. Conceptual, C. Principles, D. Procedural
Alignment Code: B3, C3, D3, B4, D4, C4, D6, D7

Set 1 (Learning Objective: Classify Sampling Procedures)

The responses options for this question on both exams are: Stratified random sample, Simple random sample, Judgmental or quota sample, and Convenience sample.

Proctored Exam: Plans for getting a sample of births for babies born in San Francisco, which has 10 districts, are given. For each plan, identify the type of sampling used.

1. Consider all 10 districts in San Francisco and randomly choose a sample within each district.
2. Determine the population of each of the ten districts of San Francisco, and then choose births so that the proportion of births in the sample is close to the ratio of the district population to the city population.
3. Make a list of hospitals and randomly sample five of them. Use as a sample the information on all of the births in those hospitals.
4. Use as the sample the information on all of the births in the biggest hospital in San Francisco.

Unproctored Exam: Plans for getting a sample of births for babies born in Georgia, which has 159 counties, are given. For each plan, identify the type of sampling used.

1. Make a list of hospitals and randomly sample seven of them. Use as a sample the information all of the births in those hospitals.
2. Use as the sample the information on all of the births in the biggest hospital in Georgia.
3. Consider all 159 Georgia counties and randomly choose a sample within each county.
4. Determine the population of each of the 159 counties. Choose the number of births so that the proportion of births in the sample is close to the ratio of the county population to the state population.

Set 2 (Learning Objective: Classify Inferential Procedures)

The response options for this question on both exams are: ANOVA, Chi-Square, Comparison of two independent means, and Comparison of two proportions.

Unproctored Exam: Statistical tasks are given for a random sample of senior swimmers who competed in 2009, 2011, and 2013. For each task, identify the appropriate inferential procedure.

1. Determine whether the average swimming time differed across the years.
2. Determine whether the average swimming time of senior swimmers differed between males and females.
3. Determine whether the proportion of 61-65 years old swimmers differed from the proportion of 66-70 years old swimmers.
4. Determine whether the proportions of male and female senior swimmers differed across the years.

Proctored Exam: Statistical tasks are given for a random sample of figure skaters who competed in 2013, 2014, and 2015. For each task, identify the appropriate inferential procedure.

1. Determine whether the average score of figure skaters differed across the years.
2. Determine whether the average score of figure skaters differed between males and females.
3. Determine whether the proportion of 14-18 years old figure skaters differed from the proportion of 19-23 years old figure skaters.
4. Determine whether the proportions of male and female figure skaters differed across the years.

Appendix E: Samples of Multiple-Choice Questions

Assigned points: 2
Cognitive process: 3. Comprehension, 4. Analysis, 7. Evaluation (Darwazeh & Branch, 2015).
Knowledge dimension: B. Conceptual, C. Principles, D. Procedural
Alignment code: B3, B4, B7, C4, D3, D7

Set 1 (Learning Objective: Interpret the CI for the Population Proportion)

Proctored Exam: Give the interpretation of the 95% confidence interval [0.048, 0.132] for the population proportion of premature births.

Select one:

Alternative

a. We are 95 % confident that the population proportion of premature births is between 0.048 and 0.132 with the margin of error of 0.042.
b. We are 95 % confident that the population proportion of premature births is between 0.042 and 0.048 with the margin of error of 0.132.
c. We are 95 % confident that the sample proportion of premature births is 0.042 with the margin of error of 0.048 or 0.132.

Unproctored Exam: Give the interpretation of the 95% confidence interval [0.153, 0.207] for the population proportion of successful students.

Select one:

Equivalent

a. We are 95 % confident that the population proportion of successful students is between 0.153 and 0.207 with the margin of error of 0.027.
b. We are 95 % confident that the population proportion of successful students is between 0.027 and 0.153 with the margin of error of 0.207.
c. We are 95 % confident that the sample proportion of successful students is 0.027 with the margin of error of 0.153 or 0.207.

Set 2 (Learning Objective: Interpret the CI for the Population Mean)

Unproctored Exam: Give the interpretation of the 95% confidence interval [65.169, 70.908] for the average age of senior swimmers.

Select one:

Alternative

a. We are 95 % confident that the population mean of senior swimmers' age is between 65.169 and 70.908 with the margin of error of 2.869.
b. We are 95 % confident that the population mean of senior swimmers' age is between 2.869 and 65.169 with the margin of error of 80.908.
c. We are 95 % confident that the sample mean of senior swimmers' age is 65.169 with the margin of error of 70.908.

Proctored Exam: Give the interpretation of the 95% confidence interval [69.648, 74.792] for the average score of figure skaters.

Select one:

a. We are 95 % confident that the population mean of figure skaters' score is between 69.648 and 74.792 with the margin of error of 2.572.
b. We are 95 % confident that the population mean of figure skaters' score is between 2.572 and 69.648 with the margin of error of 74.792.
c. We are 95 % confident that the sample mean of figure skaters' score is 69.648 with the margin of error of 2.572 or 74.792.