Journal of the National College Testing Association
2022/Volume 5/Issue 1

The National College Testing Association (NCTA) is a non-profit organization of testing professionals working in post-secondary institutions, in companies with test-related products and services, and in other professional testing venues. NCTA was organized in 2000. NCTA is dedicated to the promotion of professionalism and quality in the administration of testing services and programs, including issues relating to test administration, test accessibility, test development, test scoring, and assessment.

Editor's Note: I am honored to have served as the Editor of the *Journal of the National College Testing Association* since 2017. In this issue, Dr. Sally Carter reviews the book *Cheating in College: Why Students Do It and What Educators Can Do About It* (McCabe et al., 2012), following her leadership of a successful and widely attended NCTA book club in 2021. Dr. Gladys Bennett presents a study investigating the use of the Educational Testing Service (ETS) Criterion Online Writing Evaluation automated essay scoring (AES) with English 101 faculty. Thanks to both for furthering the professional development of NCTA and the testing profession. This volume completes my time as Editor, and I am appreciative of Dr. Steve Saladin in working with me through the transition to his Editorship. I know *JNCTA* is in great hands with Dr. Saladin and the incoming Editorial Board members.

Thank you,
Sara Rieder Bennett, Ph.D.

Email: journal@ncta-testing.org
Website: https://ncta.memberclicks.net/journal-of-the-national-college-testing-association

Journal of the National College Testing Association
2022/Volume 5/Issue 1

Book Review of *Cheating in College: Why Students Do It and What Educators Can Do About It*

SALLY CARTER, ED.D.
Southeast Missouri State University

Sally Carter is the Director of Testing Services at Southeast Missouri State University. Email: sacarter@semo.edu

Cheating in College: Why Students Do It and What Educators Can Do About It (McCabe et al., 2012) encapsulates decades of research examining academic integrity in college and university settings. Based on initial research by William J. Bowers (1964), the authors administered surveys to colleges and universities throughout the United States beginning in 1999. Their research continues today through the work of the International Center for Academic Integrity. A book discussion resulted in the following applicable takeaways for the testing industry.

The book is organized into nine chapters. After an introductory chapter, the authors discuss academic dishonesty in high schools in chapter two to provide context for their college and university data. Chapter three outlines the nine forms of cheating researched. The second section of the book, which contains chapters four through six, discusses factors that influence academic integrity. These include individual student characteristics, honor codes, peer influence, and the culture of academic integrity on campus. Chapter seven discusses faculty roles in deterring cheating. Chapter eight examines data regarding cheating in graduate programs. The final chapter provides practical advice for faculty and administrators for creating a culture of academic integrity in their schools.

Three key takeaways from *Cheating in College* resonate with the testing industry. First, methods and definitions of cheating are changing. Bowers' original work (1964) defined specific types of cheating and plagiarism instead of researching cheating as a generic term. From Bowers' list of 13 cheating behaviors, McCabe, Butterfield, and Treviño (2012) narrowed their research list to nine, combining similar items. Over the decades, self-reported incidents of cheating using the internet and collaboration have increased, while self-reported incidents of cheating on exams have decreased. What does this mean for the testing industry? First, it is important to note that the response rate for the McCabe survey has greatly declined following the move from a pen-and-paper survey to an electronic survey. The researchers feel respondents might be less honest on electronic surveys due to a perceived lack of anonymity (McCabe et al., 2012, p. 60). Further, many respondents indicated that they no longer consider behaviors such as copying answers, getting help from friends, or accessing unauthorized aides to be cheating, while people in testing most definitely would label these behaviors as cheating (p. 31). Understanding how definitions of cheating have changed from the student perspective over the past few decades informs testing personnel by helping them understand the mindset of students. While we still need to be diligent in looking for crib notes or writing on clothing, shoes, water bottles, and the like, we also need due diligence as we proctor to watch for signs of collaboration and collusion.

A second key takeaway from *Cheating in College* is the juxtaposition of the importance of academic integrity and the thoughts, feelings, and trends of students participating in the research. McCabe, Butterfield, and Treviño (2012) detail six reasons why people should care about academic integrity (p. 3). Reasons range from protecting the reputation of academia, to encouraging ethical development in young adults, to stressing the importance of understanding the pressure to cheat faced by today's students. Testing professionals could add protecting exam content and ensuring a valid, reliable, and equitable testing experience to the list. Students need to be informed of what

constitutes cheating. Examples of what schools are doing to combat cheating include rules agreements on log-in screens or signed agreements listing institutional guidelines (like the agreements test takers must sign for vendor testing), ad campaigns across campus informing students of actions constituting cheating, educating students during orientation or freshman seminar classes, and workshops to train faculty on proctoring best practices. Chapter seven outlines faculty perceptions of cheating; only 39% of faculty surveyed considered cut-and-paste plagiarism from the internet "serious cheating," with 17% labeling it as "trivial" or "not cheating" (McCabe et al., 2012, p. 132). If faculty do not understand what constitutes cheating, how can students be expected to uphold the high standards?

Finally, *Cheating in College* stresses the importance of building an institution-wide culture of academic integrity. McCabe, Butterfield, and Treviño (2012) divide their data into two groups: schools with academic honor codes and schools without academic honor codes. The classic honor code system requires students to sign a pledge stating they will adhere to integrity standards and report fellow students they see violating the honor code. In return, exams are typically not proctored. A student-led judicial system is in place to oversee code violation accusations. Schools may choose to adjust the traditional honor code formula to create a modified honor code that better serves the needs of today's students. However, the researchers question at what point dismantling the core components of a traditional honor code removes it from the classification of being an honor code. Schools with enforced honor codes have lower rates of incidents of self-reported cheating, higher faculty observation of academic integrity policy and procedure, and students who carry integrity into the workplace once out of school. Students at modified code institutions report higher rates of incidents of cheating than traditional code schools, but still a lower number of incidents of cheating compared to schools with no codes. However, it does not appear that honor codes themselves are responsible for lowering the number of incidents of academic dishonesty. Instead, "What makes the difference is that the institution's environment encourages the development and maintenance of an ethical community – what we are calling a culture of integrity" (McCabe et al., 2012, p. 167).

Cheating in College is an impactful read. Seeing decades of data confirming what testing professionals already know (e.g., stress causes people to cheat, instances of cheating are increasing, faculty and staff often feel unsupported in their work to curb cheating) serves as a rallying cry to continue pressing the issues at our institutions. And the research lives on. The International Center for Academic Integrity (ICAI) continues what is commonly referred to as the McCabe Survey. Your school can participate this year. If your school is an ICAI member, you can survey your students on their thoughts, feelings, and perceptions of academic integrity. You will receive data from your school, and your school's responses will become part of the newest decade of research. Visit https://academicintegrity.org/programs/mccabe-icai-academic-integrity-survey for more information.

REFERENCES

Bowers, W. J. (1964). *Student dishonesty and its control in college*. Bureau of Applied Social Research, Columbia University.

McCabe, D. L., Butterfield, K. D., & Treviño, L. K. (2012). *Cheating in college: Why students do it and what educators can do about it*. Johns Hopkins University Press.

Journal of the National College Testing Association
2022/Volume 5/Issue 1

Criterion Study: Assessing the Validity of the ETS Criterion Automated Essay Scoring Program for English 101 Students

GLADYS M. BENNETT, PH.D.
Norfolk State University

Author Note

Gladys M. Bennett, Ph.D., https://orcid.org/0000-0002-9226-5633

I have no conflicts of interest to disclose.

I would like to acknowledge the following, whose contributions were essential to the completion of this study and manuscript:

Dr. Gary C. Wilkens, Norfolk State University
Dr. Mamie L. Johnson, Norfolk State University
Mr. Ahmasi K. O'Daniel, Norfolk State University
Mr. Jason Norman, Norfolk State University
Mr. Damani Drew, Norfolk State University
Dr. Sara Rieder Bennett, University of Akron
Dr. Scott Debb, Norfolk State University

Correspondence concerning this article should be addressed to Gladys M. Bennett, Department of Testing Services, Norfolk State University, HB Wilson Hall, Room 133, Norfolk, VA 23504, United States: Email: gmbennett@nsu.edu.

The current study investigates the validity of the Educational Testing Service (ETS) Criterion Online Writing Evaluation automated essay scoring (AES) service for students enrolled in a first-year English course at a public state institution in southeastern Virginia. Instruments used in the study were the ETS Criterion AES service and the English and Foreign Language essay scoring rubric. Essay scores were collected from Criterion and human raters. Concurrent validity was measured using a correlational design. A significant positive relationship ($r(33) = .294$, $p = .045$) between human judgment scores and automated Criterion scores was found. The correlation approximated a medium level of correlation (Cohen, 1988) and met the standards set by the department for validity. Therefore, the Criterion was determined to be a good supplement to faculty essay scoring.

Keywords: artificial intelligence, automated essay scoring, human judgment scoring

CRITERION STUDY: ASSESSING THE VALIDITY OF THE ETS CRITERION AUTOMATED ESSAY SCORING PROGRAM FOR ENGLISH 101 STUDENTS

In recent years, the role of the Test Center Director has expanded to include not just administration and proctoring of the university and national tests, but also providing support to the university's instructional programs in selecting and validating tests for practical assessment of student learning. For those test administrators with experience and formal training in test development, it is essential to transfer that conceptual knowledge and skill to the application of testing and measurement principles to address faculty testing needs. This empirical approach can guide the faculty's development, selection, and evaluation of practical tests for improved student learning. Metrics can determine if a test is fair and unbiased, which is essential to share with university and testing communities. Fair and unbiased tests are necessary to allow faculty to accurately measure student learning and adapt instruction accordingly. We can expect to achieve optimal student learning only by accurately measuring a student's readiness to learn.

The initial pilot study of Criterion Online Writing automated essay scoring (AES) was motivated by a discussion held with the university's provost in 2017. The provost asked Testing Services to identify an artificial intelligence (AI) tool that could allow first-year English students an increased number of writing opportunities for identifying and developing writing competencies. Developing practical writing skills is a core competency required by the State Council of Higher Education and the regional accreditation commission.

Secondary objectives were to explore the possibilities of utilizing AES for incoming first-year course placement and summative assessment to reduce the time, expense, and challenge of employing multiple human judgment essay scorers. Though other AES services were considered, in the 2017 summer semester, Educational Testing Service (ETS) offered the university an opportunity to evaluate *Criterion*® Online Writing Evaluation Service (Criterion). Testing Services would use Criterion to assess English 101 student

essays. The director of Testing Services was assigned to conduct the research study.

Despite several faculty expressing interest in using Criterion to improve student writing skills, only one instructor and six students volunteered to participate in the initial pilot study. Results of correlational analysis identified statistically significant positive relationships between human judgement scores and AES scores for process analysis essays ($r(3) = .94$, $p < .05$) and career decision essays ($r(5) = .92$, $p < .01$). Following favorable anecdotal reviews from the instructor and students, Testing Services continued this study with a larger number of instructors to obtain more generalizable results (Bennett & Williams, 2017).

The current study was initiated in spring 2020 and aligned with the provost's commitment to providing faculty with innovative instructional technologies to increase student engagement and intellectual challenges. The English department identified Criterion as one of the technologies to potentially support that initiative. In 2020, the university chose "Close Reading for Effective Writing" (CREW) as its quality enhancement plan for regional accreditation. CREW describes reading and writing skills as complementary. AES was intended to supply a mechanism for students to integrate readings from core texts into their writing in support of CREW. Administrative support from the department chair was crucial to obtaining faculty support for the study.

Initially, recruiting English instructors to participate in the study was a challenge. Some instructors expressed concern that the AES scoring software would replace instructor human judgment scoring and threaten their job security. To reduce their fears, the Director of Testing Services spoke at an English and Foreign Language department faculty meeting to ensure faculty understood that Criterion's purpose was to supplement, not replace, their essay grading efforts. Another challenge was faculty acceptance of AES as a practical and effective alternative to human judgment scoring. This concern was addressed by supplying evidence that, although AES was limited in evaluating such writing elements as content, tone, and creativity, it effectively assessed fundamental skills of grammar, word usage, mechanics, style, and organization. Most faculty accepted that AES could be beneficial as a supplemental resource for essay scoring.

Many instructors agreed they would like to assign more writing projects; however, the time to grade the current assignments was challenging. Instructors were informed of the potential of AES to free up a large amount of instructor time used to develop fundamental writing skills so they could instead focus on developing higher-order writing skills.

The e-rater® automated scoring engine automates Criterion. The service uses natural language processing (NLP) technology, a branch of AI, to evaluate several aspects of writing proficiency, including grammar, usage, mechanics, and development. "The features of e-rater have been periodically updated and enhanced to improve engine performance" (Chen et al., 2017, p. 2). Beyond scoring, Criterion helps students plan, write, and revise their essays. Students are provided immediate feedback with increased opportunities to practice writing and improve the quality of their writing.

Purpose of the Present Study

The university's challenge was to identify a practical solution to reduce the time,

expense, and challenge of employing multiple human judgment scorers to read and score entering first-year student essays for use in English course placement and student assessment. Utilizing AI, Criterion would provide English 101 students more opportunities to develop writing competencies without requiring additional instructor time to evaluate and grade essays.

Assessments are more meaningful when they are locally validated. Therefore, the focus of this study was to validate Criterion as an appropriate measure to assess English 101 student writing. The English department's criteria for writing competency were used as a basis for comparison. This study determined if student essay scores on Criterion, a well-established automated scoring program in higher education, were significantly correlated with human judgment essay scores submitted by English 101 instructors. The human judgment scorers were asked to address the following writing quality features that are scored by Criterion: grammar errors (e.g., subject-verb agreement), usage (e.g., preposition selection), mechanics (e.g., capitalization), style (e.g., repetitious word use), discourse structure (e.g., presence of a thesis statement, main points), vocabulary usage (e.g., relative sophistication of vocabulary), sentence variety, source use, and discourse coherence quality.

A positive correlation would indicate that the two measures were similar, and that Criterion could be an alternative for student essay scoring.

Review of the Literature

Writing is an essential life skill that enhances the ability to communicate effectively and critically assess the writing of others. It expands knowledge in various subject areas for success in school, employment, and personal development. "From a faculty member's perspective, writing well entails more than adhering to writing conventions. Writing also encompasses creative inspiration, problem-solving, reflection, and revision, resulting in a completed manuscript" (DeFazio et al., 2010, p. 34).

In higher education institutions with limited financial resources and increasing class sizes for English instructors, AES is a means to capitalize on the use of AI as a supplement to human resources to provide students with expanded writing opportunities to develop essential writing skills.

There is consensus in higher education and employment that writing is essential for developing practical communication skills, thinking, and creativity. These skills are required for success in college, the workplace, and society. The National Commission on Writing for America’s Families, Schools, and Colleges (2004) cited results of a 2003 benchmark report which inferred that because it is time-intensive for students and instructors, writing is the "most neglected" (p. 22) of the three fundamental skills taught in schools and colleges.

The extensive time required to evaluate and grade writing assignments is a significant constraint on providing college students with sufficient practice in developing writing skills. Typically, multiple writing assessment methods are used to assess and establish first-year student writing skills, including standardized tests, writing samples, course grades, and portfolio assessments. In recent years, there has been a trend toward writing in digital environments and increasing automated scoring and feedback availability. In a 2013 position paper, the National Council of Teachers of English acknowledged the

potential financial savings associated with computerized writing assessment. However, they reported a detrimental impact of automated assessments, neglecting the complex and varied types of writing found in higher education. Nevertheless, the trend toward the use of AES has continued.

Klobucar and colleagues (2011) reported more than twelve different automated essay evaluation systems designed to predict human scores. Rudner and Gagne (2001) recognized the three major mechanical scoring programs as Project Essay Grade, Intelligent Essay Assessor, and e-rater, the engine that drives the Criterion Online Writing Evaluation Service.

The process for developing the AI automated essay scoring programs is highly dependent on human subject scoring. It begins with collecting a representative sample of student writings in response to a given prompt. The essays are then hand-scored by qualified graders. The AI team develops the model for scoring new student essays using a statistical method derived from the human scores.

Ellis B. Page was considered the father of AES and designed Project Essay Grade in 1966 (Wang & Brown, 2007). Page's objective was to "make the analysis of language economically efficient and educationally sound" (Page, 1968, p. 211). He considered the computer a "natural agent in the analysis of language" (p. 211), provided that the rules for evaluating student writing are clearly specified.

Page (1968) focused his research on content and style. At least four independent experts judged sample student essays for overall quality. Hypotheses were then generated for variables that might be associated with these judgements. Page wrote computer programs to measure these variables. He later identified five crucial traits: ideas, organization, style, mechanics, and creativity. Page's research indicated that computer scores on the crucial traits were well correlated with those of human judges. Page's identified traits continue to be considered essential to effective essay writing.

Advancements were made in AES in the 1990s, and devices such as ETS's e-rater were developed to include advanced writing elements such as "syntactic variety, topic content, and organization of ideas" (Wang and Brown, 2007, p.8). The e-rater was initially developed to grade the Analytic Writing Assessment section of the Graduate Management Admissions Test (GMAT) and uses NLP techniques to predict human raters' scores best. Pearson Knowledge Technologies developed another advanced AES, the Intelligent Essay Assessor (IEA). IEA uses the Latent Semantic Analysis approach, which proposes that words close in meaning appear in similar text sections. IEA provides an evaluation of essay content, style, and mechanics. IntelliMetric was developed by Vantage Learning and reported to have blended AI with NLP and statistical technologies, allowing for identifying characteristics that human scorers value and those they find deficient (Wang & Brown, 2007).

Wood (2013) suggested that the ideal use of AES would be in conjunction with human readers to manage increased writing demands and "assure better quality scoring of student writing" (para. 3). A second evaluation would increase the reliability of scores. Foltz and colleagues (1999) concluded that human scoring is still essential for scoring content and creativity of students' writing because computers do not read.

Though there are detractors of AES (National Council of Teachers of English, 2013; Perelman, 2013), research has shown

(Dikli, 2006; Giles, 2011; Klobucar et al., 2011) that AES can be an asset for scoring grammatical and structural writing elements. Lewis (2013) cautioned that there had been little independent testing of commercial AES systems, and the software developers provide most data regarding correlations with human scorers.

Validity assessments of AES are inconclusive. Shermis and colleagues (2002) examined the validity of the Project Essay Grade and found the computer ratings to be at least as valid as pairs of human judges (Wang & Brown, 2007). A validation study by Landauer (2010) indicated that IEA correlated with human graders as reasonably as they correlated with each other; however, McGee (2006) concluded that IEA did not live up to its claims of grasping the meaning of the text and was inadequate for scoring student essays. According to Powers and colleagues (2000), e-rater demonstrated significant "but modest" (Abstract) agreement with human raters in scoring Graduate Record Exam (GRE) Writing Assessments and suggested that the two methods of scoring reflected similar elements of writing proficiency. A research report on mean differences between e-rater automated scoring and human scoring of the GRE showed that, though e-rater was trained to maximize the prediction of human scores, there were notable differences in scoring for some demographic subgroups. Also, e-rater was less severe on language errors, overvalued organization and development, and occasionally undervalued content relative to human raters (Ramineni & Williamson, 2018).

Criterion appears to be highly effective in scoring grammar, usage, mechanics, and style but is limited in grading content. Until further advancements in AES technology, that element of the students' writing will continue to require human judgment scoring. More research is needed to identify additional factors explaining differences in AES and human scoring.

For the purposes of this study, we decided to assess the viability of Criterion independently as an alternative scoring method for college student essays. The small sample size and lack of variability in participants' ethnicities prevented data analysis by student demographic groups.

Hypothesis

Based on the results of previous research and results of the preliminary Criterion pilot study conducted by the researcher in the summer of 2017, the hypothesis was that there would be a statistically significant positive correlation between the human judgment scores and the automated Criterion scores. The null hypothesis was that there would not be a statistically significant difference between the two types of scoring.

METHOD

Participants

The principal investigator recruited 32 students, age 17 or older, from English 101 courses and obtained the students' in-person or electronic consent to participate in the study. Students were informed of the purpose, risks, potential benefits of the study, and incentives to participate. Students were told that their participation was voluntary and could withdraw from the study without penalty (Appendices A and B).

Sampling Method

Non-probability sampling was used to select study participants. All departmental faculty teaching English 101 courses were invited to

participate in the study. Four faculty chose to participate in the study. All students enrolled in the four English 101 courses were asked to participate. Of the 32 student participants, 50% identified as female and 50% identified as male. Ethnicities reported were 97% Black/Non-Hispanic and 3% American Indian/Alaskan Native.

Materials/Measures

Criterion provided AES in five key categories — Grammar, Usage, Mechanics, Style, and Organization & Development. The ETS service is powered by a patented *e-rater*® scoring engine that provides annotated diagnostic feedback and holistic scoring based on level-specific models built from essays pre-scored by ETS-trained readers. The holistic score is a single numerical score given to the essay as a whole and reflects the overall quality of the writing. Scores range from one to six or one to four, with one being the lowest. For purposes of this study, Criterion and human scores were converted to percentage scores for comparative analyses.

The English faculty used the department of English and Foreign Languages Essay Scoring Rubric (Norfolk State University, 2020) to score the essays. Faculty ratings were used to represent human judgment in this study. The rubric guided scoring categories of organization, development & analysis, sentence structure, grammar, diction, and mechanics. Scores ranged from one to five, with one being the lowest possible score.

Procedure

During the spring 2020 and fall 2020 semesters, 32 students in English 101 wrote and submitted essays in word format to their professor and copied them to the Criterion automated essay scoring program. The students were informed that the Criterion scores would not be averaged with their final grade but counted as extra credit points.

The study was halted following the spring 2020 transition from in-person to online course instruction due to the COVID-19 pandemic. It was continued in fall 2020 with a modified methodology to allow for online consent (Appendix B). As an incentive to participate, the first 50 students' names were entered in a raffle with a chance to win one of six $25 visa gift cards. This information was communicated to students in an electronic flyer (Appendix C).

During the spring 2020 semester, students were requested to submit three essays; however, most participants discontinued submissions following the COVID-19-related suspension of in-person classes. Based on the English 101 modified calendar for fall 2020, the students were requested to submit only two essays. Each essay was graded by the English instructor using the Department of English and Foreign Languages (ENFL) essay scoring rubric (Appendix D) and afterward scored by Criterion. The alignment of grading criteria for ENFL scoring and Criterion scoring is shown in Appendix E. A statistical test of concurrent criterion-related validity was performed between the paired essay scores to determine if students scored similarly on both scoring methods. The Pearson correlation was used to compare the two essay scores and determine a statistically significant correlation.

Following the instructors' respective English 101 syllabi, students submitted essays for two of four essay types, which included description, classification, process analysis, and compare/contrast. Student essays were scored by their assigned English instructor before the essays were submitted

to Criterion. Instructors were not given access to the students' Criterion scores during the study to avoid experimenter bias. This approach meant that the scores entered by faculty were not influenced by the Criterion scores and were independent of them.

Data Analysis

The Pearson correlation was used to assess concurrent criterion-related validity and determine if there was a statistically significant correlation between the Criterion automated essay scoring and faculty human judgment scoring (Table 1). A one-way analysis of variance (ANOVA) was performed to compare the means between groups and determine statistically significant differences (Table 2).

Table 1

Descriptive Statistics and Correlations for Faculty and Criterion Percentage Scores

Variable	*n*	*M*	*SD*	1	2
1. Faculty % Score	45	79.733	14.298	-	.294*
2. Criterion % Score	34	78.633	11.438	.294*	-

*$p < .05$

Table 2

One-way Analysis of Variance (ANOVA) for Group Means

		SS	*df*	*MS*	*F*	*p*
Instructor % Score	Between Groups	4106.686	3	1368.895	11.482*	.000
	Within Groups	4888.114	41	119.222		
	Total	8994.800	44			
Criterion % Score	Between Groups	603.725	3	201.242	1.613	.204
	Within Groups	4367.785	35	124.794		
	Total	4971.510	38			

$p < .001$

RESULTS

Results of the Pearson correlation indicated a positive and statistically significant relationship between human judgment scores and automated Criterion scores ($r(33) = .29$, $p = .045$), one-tailed test). Results represent a moderate correlation based on Cohen's (1988) effect size recommendations for *r* (small = .10, medium = .30, and large = .50). Correlations are presented in Table 1.

During exploratory analyses, a one-way ANOVA to identify differences between instructor group means found statistically

significant differences for human scores ($F(3, 41) = 11.48$, $p < .001$). Results of the one-way ANOVA did not indicate a statistically significant difference between instructor group means for the automated Criterion scores ($F(3, 35) = 1.61$, $p = .20$). Group means are compared in Table 2.

Further analysis of group means with Scheffe's post hoc analysis indicated scores given by Faculty 1 were significantly lower than those of all other faculty ($p < .05$). Faculty 4 scores were significantly higher than those of Faculty 1 ($p < .01$) and Faculty 3 ($p < .05$). The standard deviation in Faculty 4 scores was zero, indicating no variability in scoring.

DISCUSSION

There was a statistically significant positive relationship between human judgment scores and automated Criterion scores, indicating concurrent validity. We, therefore, rejected the null hypothesis. Criterion appears to be a valid measure for scoring student essays. Criterion can help evaluate organization & development, grammar, usage, mechanics, and style. As observed in an earlier pilot study, we acknowledge the limitations of Criterion for assessing the content, organization, tone, and creativity of student essays. However, the use of Criterion can reduce faculty time spent on scoring writing mechanics and increase the amount of instructor time devoted to developing essential higher-order writing skills.

One-way ANOVA results indicated statistically significant differences between instructor group means with human subject scoring, suggesting either difference in faculty interpretation of scoring criteria or possibly differences in the standards used for essay scoring. Ramineni and Williamson (2018) referenced research studies that identify a range of problems and concerns with human scoring of essays, "including halo effects, fatigue, tendency to overlook details, and problems with consistency of scoring across time" (p. 6). Comparatively, one-way ANOVA results for automated Criterion mean scores did not indicate statistically significant differences between instructor groups; this suggests that AES provided a more objective application of the scoring criteria than human subject scoring.

Implications

The findings of this study support the use of Criterion as a supplement to faculty scoring. Criterion can increase students' opportunities for writing practice and evaluation with immediate feedback, which will allow for continual revision and improvement of their writing. Results also suggest the need to apply more standardized scoring criteria for human subject scoring. A rubric such as the standardized Association of American Colleges & Universities' written communication VALUE rubric (2009) may serve as a guide for revising the existing rubric used by English faculty. English faculty should be trained to use rubrics for more objective grading effectively. The use of Criterion combined with the effective use of standardized scoring rubrics by faculty should result in improved student writing.

Limitations

This study allowed for an objective analysis of Criterion to determine the efficacy of the Criterion automated essay scoring program for writing courses. However, because this study could not control for multiple components of English 101 final grades, such as attendance, homework, quizzes, and projects, the study did not attempt to determine whether Criterion alone was a good predictor of students' final grades in

the English 101 courses. Another limitation of the study was the inability to obtain multiple faculty ratings for each submitted essay; thereby, assessment of inter-rater reliability was not possible.

The researcher believed that participation in the current study was negatively impacted when the study was halted three months into the spring 2020 semester due to the COVID-19 pandemic. The pandemic subsequently reduced access to participants who were transitioned from in-person to online instruction. As a result, the number of student participants and essays submitted were less than intended and limited the generalizability of the results.

The post hoc analysis indicated highly variable scores for the faculty raters. Some faculty had minimal variation in their scores and did not seem to be applying the rubric, whereas other faculty seemed to be following the rubric very carefully. Procedures to ensure accountability among the graders to adhere to the rubric were not evident.

Future Research

Continued research should be conducted to collect and analyze English course completion data to compare the success rates of students who use the Criterion service with those who choose not to use the Criterion service. Additional AES programs should be evaluated to determine which is most in agreement with human judgments of student essays. Results of these studies should guide decision-making on adopting the most appropriate AES program(s) for the university.

Further research is warranted to investigate the efficacy of standardized rubrics for increasing comparability in human subject scoring of English writing assignments. Faculty subjectivity cannot be eliminated from the scoring process; however, the use of standardized rubrics can improve the consistency of faculty grading. Future research designs should include selecting faculty graders other than the instructor of record to provide more than one human judgment score for each submitted essay and allow for assessment of inter-rater reliability in human essay scoring.

Conclusion

The findings of this study suggest that Criterion automated essay scoring can be used in introductory English courses as a supplement to faculty essay scoring. Criterion was shown to be more consistent than faculty grading. While benefiting students with increased opportunities for practice and development of writing skills, Criterion can reduce the faculty time required for grading essays. Such time savings can allow faculty to focus on developing students' higher order writing skills, resulting in improved student writing.

REFERENCES

Association of American Colleges and Universities. (2009). *Written communication VALUE rubric*. https://www.aacu.org/value/rubrics/written-communication

Bennett, G., & Williams, M. (2017). *NSU Criterion pilot study: Assessing the validity of the ETS Criterion automated essay scoring program for NSU English 101 students* [Unpublished manuscript]. Norfolk State University.

Chen, J., Zhang, M., & Bejar, I. I. (2017). An investigation of the e-rater® automated scoring engine's grammar, usage, mechanics, and style microfeatures and their aggregation model. *ETS Research Report Series, 2017*(1), 1-14. https://doi.org/10.1002/ets2.12131

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Lawrence Erlbaum Associates.

DeFazio, J., Jones, J., Tennant, F., & Hook, S. A. (2010). Academic literacy: The importance and impact of writing across the curriculum – A case study. *Journal of the Scholarship of Teaching and Learning, 10*(2), 34-47. https://files.eric.ed.gov/fulltext/EJ890711.pdf

Dikli, S. (2006). An overview of automated scoring of essays. (2006). *Journal of Technology, Learning, and Assessment, 5*(1). https://ejournals.bc.edu/index.php/jtla/article/view/1640

Educational Testing Service. (n.d.) *The Criterion® online writing evaluation service*. http://www.ets.org/criterion.

Foltz, P. W., Lahan, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. In B. Collis & R. Oliver (Eds.), *Proceedings of ED-MEDIA 1999 - World Conference on Educational Multimedia, Hypermedia & Telecommunications* (pp. 939-944). Seattle, WA, United States. Association for the Advancement of Computing in Education. https://www.learntechlib.org/p/6607/

Giles, J. (2011, August 31). *Automated marking takes teachers out of the loop*. New Scientist. https://www.newscientist.com/article/mg21128285-200-automated-marking-takes-teachers-out-of-the-loop/

Klobucar, A., Dean, P., Elliot, N., Ramineni, C., Deess, P., & Rudniy, A. (2011). Automated essay scoring and the search for valid writing assessment. In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 103-120). Parlor Press, LLC.

Landauer, T. K. (2010). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10*(3). https://doi.org/10.1080/0969594032000148154

Lewis, J. K. (2013). *Ethical implementation of an automated essay scoring (AES) system: A case study of student and instructor use, satisfaction, and perceptions of AES in a business law course*. Salve Regina University Digital Commons. http://dx.doi.org/10.2139/ssrn.2684803

McGee, T. (2006). Taking a sin on the intelligent essay assessor. In P. Freitag & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 79-92). USU Press Publications. https://digitalcommons.usu.edu/usupress_pubs/

National Council of Teachers of English. (2013, April 20). *NCTE position statement on machine scoring*. http://www.ncte.org/positions/statements/machine_scoring

Norfolk State University. (2020). *Department of English and Foreign Languages essay scoring rubric*.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education, 14*(2), 210-225. https://www.jstor.org/stable/3442515

Perelman, L. C. (2013). Critique of Mark D. Shermis & Ben Hamner, 'contrasting state-of-the-art automated scoring of essays: analysis'. *Journal of Writing Assessment, 6*(1), 1-10. https://escholarship.org/uc/item/7qh108bw

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000, June). *Comparing the validity of automated and human essay scoring* (GRE Board Research Report No. 98-08aR, ETS Research Report 00-10). Educational Testing Services. https://www.ets.org/Media/Research/pdf/RR-00-10.pdf

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® general test. *ETS Research Report Series, 2018*(1), 1-31. https://doi.org/10.1002/ets2.12192

Rudner, L., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation, 7*(26). http://ericae.net/pare/42~getvn.html

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62*(1), 5-18. https://doi.org/10.1177/0013164402062001O1

The National Commission on Writing for America's Families, Schools, and Colleges. (2004, September). *Writing: A ticket to work... or a ticket out: A survey of business leaders*. College Board. https://archive.nwp.org/cs/public/print/resource/2540

Wang, J. & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment, 6*(2), 1-29. https://eric.ed.gov/?id=EJ838612

Wood, J. (2013). *Teacher and automated essay scoring (AES)... a winning combination?* Retrieved July 17, 2017, from https://www.nwea.org/blog/2013/teacher-and-automated-essay-scoring-aes-a-winning-combination/

APPENDIX A

Criterion Consent Form

You are being asked to take part in a research study of automated essay scoring for XXX English 101 students. We are asking you to take part because you are currently enrolled in an English 101 course. Please read this form carefully and ask any questions you may have before agreeing to take part in the study.

What the study is about: The purpose of this study is to learn how the ETS Criterion automated essay scoring program compares with human judgement scoring by an English instructor. You must be enrolled in English 101 to take part in this study.
What we will ask you to do: If you agree to participate in this study, you will be asked to write three essays that will be scored both by Criterion and an English instructor.

Risks and benefits: I do not anticipate any risks to you participating in this study other than those encountered in day-to-day life. Though there are no direct benefits to you, your participation in this study will help us to identify efficacious essay scoring measures that allow for immediate feedback and an increased number of writing opportunities that will help to develop writing competency.

Compensation: You may earn extra credit if you are taking a class that offers credit for research studies. The class instructor will assign credit according to class policy.

Your answers will be confidential. The records of this study will be kept private. In any sort of report that we make public we will not include any information that will make it possible to identify you. Research records will be kept in a locked file; only the researchers will have access to the records.

Taking part is voluntary: Taking part in this study is completely voluntary. If you decide to take part, you are free to withdraw at any time.

If you have questions: The researchers conducting this study are xxxxxxxxx. Please ask any questions you have now. If you have questions later, you may contact xxxxxxxxxxxxxx. If you have any questions or concerns regarding your rights as a subject in this study, you may contact the Institutional Review Board (IRB) Chairperson, xxxxxxxxxxxxxxxxx . You may also access the IRB website at xxxxx.You will be given a copy of this form to keep for your records.

Statement of Consent: I have read the above information and have received answers to any questions I asked. I consent to take part in the study.

Your Signature _____ Date _____

Your Name (printed) _____

Signature of person obtaining consent _____ Date _____

Printed name of person obtaining consent _____ Date _____

This consent form will be kept by the researcher for at least three years beyond the end of the study.

APPENDIX B

Revised Criterion Consent Form

You are being asked to take part in a research study of automated essay scoring for XXX English 101 students. We are asking you to take part because you are currently enrolled in an English 101 course. Please read this form carefully and ask any questions you may have before agreeing to take part in the study.

What the study is about: The purpose of this study is to learn how the ETS Criterion automated essay scoring program compares with human judgement scoring by an English instructor. You must be enrolled in English 101 to take part in this study.

What we will ask you to do: If you agree to participate in this study, you will be asked to write three essays that will be scored both by Criterion and an English instructor.

Risks and benefits: I do not anticipate any risks to you participating in this study other than those encountered in day-to-day life. Though there are no direct benefits to you, your participation in this study will help us to identify efficacious essay scoring measures that allow for immediate feedback and an increased number of writing opportunities that will help to develop writing competency.

Compensation: You may earn extra credit if you are taking a class that offers credit for research studies. The class instructor will assign credit according to class policy.

Your answers will be confidential. The records of this study will be kept private. In any sort of report that we make public we will not include any information that will make it possible to identify you. Research records will be kept in a locked file; only the researchers will have access to the records.

Taking part is voluntary: Taking part in this study is completely voluntary. If you decide to take part, you are free to withdraw at any time.

If you have questions: The principal investigator for this study is xxxx. If you have questions, you may contact xxxx. If you have any questions or concerns regarding your rights as a subject in this study, you may contact the Institutional Review Board (IRB) Chairperson, xxxx. You may also access the IRB website at xxxx. You will be given a copy of this form to keep for your records.

Statement of Consent: I have read the above information and have received answers to any questions I asked. I consent to take part in the study.

Your typed name confirms consent
Your Signature _____ Date _____
Your Name (printed) _____
Signature of person obtaining consent _____ Date _____
Printed name of person obtaining consent_____ Date _____

This consent form will be kept by the researcher for at least three years beyond the end of the study.

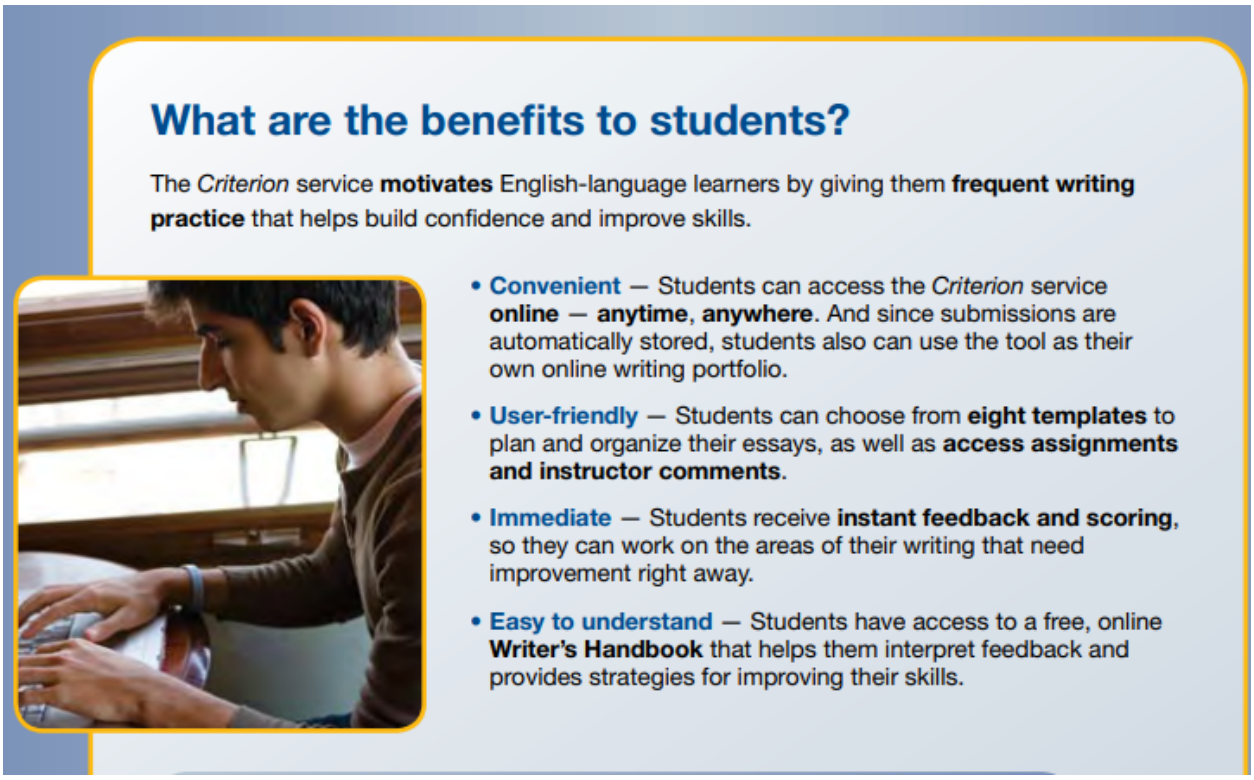
APPENDIX C

Electronic Recruitment Flyer to be Emailed to English 101 Students

ETS Criterion

JOIN US!

As an English 101 student, you are invited to participate in a study of automated essay scoring using artificial intelligence. There is no financial cost for participation, and you will be able to use this service for the remainder of the year for any of your courses! You simply copy and paste your essay to receive an immediate score and feedback.



If you are interested in participating in this study, please xxxx no later than October 21, 2020.

If you are one of the first 50 volunteers, you will be entered into a raffle and have a chance to win one of two $25.00 visa gift cards!

APPENDIX D

Department of English and Foreign Languages Essay Scoring Rubric

Department of English and Foreign Languages					
	100-90	**89-80**	**79-70**	**69-60**	**59-50**
EWC Scoring Rubric	**5 (Superior Competency)**	**4 (Above Average Competency)**	**3 (Competency)**	**2 (Developing Competency)**	**1 (Incompetence)**
Organization - Appropriate use of essay structure (Introduction, Thesis statement, Body paragraphs, Conclusion, Transitional devices, etc.)	Clearly-stated, sophisticated thesis directly addresses the prompt; Introduction establishes the content and purpose; Conclusion effectively recounts and summarizes arguments; Body paragraphs include main points discussed separately and in detail; Effective use of thoughtful transitions that connect ideas.	Clearly-stated thesis addresses the prompt; Introduction begins to establish a foundation for the content and purpose; Conclusion summarizes arguments; Body paragraphs are sound and reinforce structure; Transitions connect ideas.	Generalized thesis addresses the prompt; Simple, but recognizable introduction and conclusion; Adequate incorporation of support for thesis in body paragraphs, though they may obtain some extraneous information; Transitions may be mechanical, but foster coherence.	Thesis is vague or implied, not clear or specific, may simply broach prompt; Introduction and conclusion do not establish purpose or summarize arguments; Body paragraphs are poorly organized, ideas are strung together haphazardly; Ineffective transitions.	No clear or implied thesis statement; No clear introduction or conclusion; Paragraphing is missing, irregular or so frequent that it has no relationship to the essay; transitions are confusing or absent; Organizational problems make the essay near impossible to understand.
Development & Analysis - Appropriate use of central ideas and concrete details that support the thesis and prompt	Arguments effectively address all aspects of the prompt; Relevant, quality details enrich the central theme; Shows clear insight on the part of the writer.	Details are present and support arguments; Arguments are clear and illustrate some awareness of the complexities of the issue being discussed.	Development is basic, ideas are reasonably clear, though they do not help flesh out some of the main arguments presented; Arguments on topic, but may not demonstrate in-depth understanding.	Details may be too broad, narrow or inappropriate; Arguments are unclear or supporting evidence is insufficient, often unnecessarily repetitious.	Supporting information is limited, unclear or not present at all; Thoughts are disconnected and have no discernable point; Essay length is not adequate for development.

Sentence Structure - Appropriate use of the construction of complete, complex sentences	Complete sentences are well-built with complex and varied structure; Little to no sentence structure errors such as fragments, run-ons etc.	Sequencing is logical and effective, some sentence variety and use of complex sentence forms; Very few fragments, run-ons etc.	Sequencing shows logic, some sentence variety; Sentences are routine, but effective; A few fragments, run-ons, etc., but not to the point of distraction.	Very little sentence variety, most are structured the same way; Some are awkward, others are fragments, run-ons, etc.	Sequencing is random, most phrases are not sentences at all; Endless conjunctions or a complete lack thereof, which causes mass confusion.
Grammar, Diction & Mechanics- Appropriate use of the conventions of grammar such as tense, POV, as well as language usage, punct., spelling, capitalization, etc.	Little to no grammatical errors (i.e. subject/verb agreement, tense, POV) used effectively and coherently throughout the essay; Language choices enhance meaning and clarify understanding in a precise, interesting way; Near perfect execution of internal and external punctuation, spelling and capitalization (**1-3 errors**).	A few grammatical errors, but grammar is correctly applied; Attempt at use of varied and advanced language that enhances arguments; Very few external punctuation and a few internal (i.e. comma, semi-colon, etc.) errors; Very few spelling and capitalization errors (**3-5 errors**).	Problems with grammar are not serious enough to distort meaning, but may not be correctly applied in each instance; Attempts at colorful language apparent, but diction sometimes reaches beyond the scope of the argument; Punctuation sometimes missing or wrong; Some spelling and capitalization errors (**5-10 errors**).	Numerous grammatical errors that distort meaning in some instances; Language often used in odd ways; Jargon or clichés distract or mislead, redundancy is distracting; Many external and internal punctuation errors as well as numerous errors in spelling and capitalization (**10-15 errors**).	Frequent grammatical errors distort meaning and hinder communication; Little to no variation in word choice, language is used incorrectly and seriously impairs understanding; Gross errors in punctuation, spelling, and capitalization that hinder meaning and understanding. (**15+ errors**).

APPENDIX E

Crosswalk Indicating the Alignment of Grading Criteria for Human Judgment Scoring and ETS Criterion Automated Essay Scoring

Alignment of ENFL Human Scorer Grading Rubric Criteria with Criterion Traits Criteria

Grade Criteria		
ENFL	Criterion Traits	Examples
Organization - Appropriate use of essay structure (Introduction, Thesis statement, Body paragraphs, Conclusion, Transitional devices, etc.)	Organization & Development	• discourse structure (e.g., presence of a thesis statement, main points)
Development & Analysis - Appropriate use of central ideas and concrete details that support the thesis and prompt	Style	• sentence variety • source use • discourse coherence quality
Sentence Structure - Appropriate use of the construction of complete, complex sentences	Usage	• style (e.g., repetitious word use) • vocabulary usage (e.g., relative sophistication of vocabulary)
Grammar, Diction & Mechanics - Appropriate use of the conventions of grammar such as tense, POV, as well as language usage, punctuation., spelling, capitalization, etc.	Grammar, Mechanics, and Usage	• errors in grammar (e.g., subject-verb agreement) • usage (e.g., preposition selection) • mechanics (e.g., capitalization)

Journal of the National College Testing Association
2022/Volume 5/Issue 1

Comparing Performance on Entry Assessments by Proctoring Modality During the COVID-19 Pandemic

CINDY L. JAMES, PH.D.
Thompson Rivers University

Author Note

I would like to thank the TRU Assessment Centre staff, Tasha Baker, Linda Giddens, and Christene Hubbard, for coming back to campus during the pandemic so we could continue to provide professional testing services to our students and clients. It is an honor and pleasure to work with all of you. I am also grateful for the discussions with Dr. Dustin Van Gerven, Camosun College, about online testing which helped me frame this study.

Dr. Cindy L. James is the Principal Coordinator of the Assessment Centre & Chair of the Counselling, Academic Support and Assessment department at Thompson Rivers University. She is also the President of the Canadian Higher Education Testing Association (CHETA). Email: cjames@tru.ca

The COVID-19 pandemic precipitated an expansion and surge in online proctoring services, especially in higher education settings. Online proctoring is designed to be as effective as in-person proctoring, but research validating this assumption is limited. This study addresses the paucity of information by comparing the impact of in-person and online proctoring on entry assessment scores during the worldwide pandemic. The results revealed no significant differences in scores for three English and two mathematics tests by proctor type but did reveal a significant score difference for an advanced mathematics test. For this test, the students proctored online scored higher than students proctored in-person. Factors related to online proctoring, the pandemic, and the design of the test provide some insight into these findings.

Keywords: online proctoring, in-person proctoring, ACCUPLACER®, higher education

COMPARING PERFORMANCE ON ENTRY ASSESSMENTS BY PROCTORING MODALITY DURING THE COVID-19 PANDEMIC

The purpose of proctoring is to ensure tests are administered according to standardized rules and procedures in order to maintain the academic integrity of testing (American Educational Research Association et al., 2014). In higher education, the gold standard of proctoring involves in-person proctoring by testing professionals working in official test centers. With the expansion of distance and online learning and the advancement in technology, the alternate form of online proctoring became viable. Online proctoring is similar to in-person proctoring, but since the proctor is not in the same location as the candidate, the proctor must rely upon technology such as web cameras, microphones, and monitoring software to observe and supervise the test session.

Over the past two decades, the growth of online proctoring has been steady. Then, the COVID-19 pandemic arrived in 2020. This worldwide pandemic completely disrupted higher education, with institutions closing their campuses to avoid transmission of the potentially deadly disease and moving to virtual services on a temporary basis. The impact on testing services, especially test centers, was profound. Many testing centers were closed for several months. Some stayed closed for over a year, with testing staff working virtually. Test centers that stayed open or reopened had to limit testing volume and follow pandemic protocols. For the former, staff could no longer proctor in-person, so to resume testing services, most transitioned to online proctoring, utilizing either the commercial services available or assuming the role themselves. As for the test centers that reopened or remained open, in some cases the limited capacity for in-person proctoring was not sufficient to meet demand, so online proctoring was employed to supplement in-person proctoring.

Various vendors offer online proctoring services, providing both a platform and trained proctors to supervise exams for a fee (e.g., ProctorU, Examity®, Proctortack, OnVue, Proctorio, ProProctorTM). Some institutions in higher education also offer online proctoring utilizing their learning platforms, testing staff, and faculty. Information about the prevalence of online proctoring in higher

education is limited, but the reports that do exist provide some stunning data. For instance, in November 2020, an online search of North American universities and colleges' webpages found that 63% mentioned proctoring software (Kimmons & Veletsianos, 2021). Although this survey does not provide any insights into the degree to which the tools are used, it does indicate that online proctoring tools and services are commonplace in higher education, albeit more so in the United States (65.8%) than in Canada (39.2%; Kimmons & Veletsianos, 2021). Reviewing websites of specific online proctoring vendors also highlights the dramatic rise of these services. For instance, Proctorio, established in 2013, remotely proctored over 425,000 exams in 2016; by 2020, Proctorio proctored over 20 million exams at more than 2,000 institutions worldwide (Proctorio, 2022). Examity®, also founded in 2013, now provides online proctoring services for over 500 academic institutions and testing organizations, testing millions of candidates worldwide (Examity, 2022c). Similarly, ProctorU started providing online proctoring services to higher education institutions in 2008 and now proctors approximately five million exams globally on an annual basis (MeazureLearning, 2022).

This expansion and recent surge in online proctoring services necessitates the evaluation of these various services. The primary research question asks, "Is online proctoring as effective as in-person proctoring?" To ensure the integrity of test sessions, it is critical that testing conditions for all candidates are standardized. Although both proctoring modalities provide live proctors who follow standard protocols, the testing environments differ. In-person testing usually takes place in testing centers that are designed to minimize security risks and distractions. Online testing often takes place inside the candidate's home, which is not an ideal setting for testing. Furthermore, testing centers typically are equipped with modern computers and high-speed internet to ensure seamless testing. In comparison, while the online proctor may be using the latest technology, the candidate may not. Their laptop or home computer may not have the best operating system to support constant monitoring, and their wireless connection or internet service could be unreliable. So, candidates' access to and level of comfort with technology could impact performance by proctor modality, as could the appositeness of the testing space.

Besides the environment, differences in how candidates are monitored by proctor type may also influence performance. Online proctoring involves having a total stranger peer into a candidate's private space (e.g., bedroom) for the duration of the exam. While this is happening, the candidate is required to stay focused on the screen, as any deviancies such as leaving to take a restroom break or even looking away from the screen can trigger red flags in the system. This type of interaction could increase the stress and anxiety level for some candidates and potentially hinder performance. In comparison, in-person testing takes place in a common location (e.g., test center) with proctors usually monitoring students from a comfortable distance and allowing for a bit more flexibility in terms of movement and breaks. However, it is important to acknowledge that during the pandemic, in-person proctoring may have also been stressful for some candidates, given the risk of exposure to COVID-19 and the strict protocols (e.g., masking, social distancing, sanitizing) that had to be followed.

Given these differences and their impacts, it is vital to investigate how in-person proctoring compares to online proctoring in terms of outcomes. The present study was designed to do just that by exploring the effects of proctoring modality (in-person proctoring and online proctoring) on entry assessment English and mathematics test scores during the COVID-19 pandemic.

LITERATURE REVIEW

Within the literature, many terms are used to describe proctoring. For instance, in-person proctoring is also referred to as onsite live proctoring. Online proctoring is commonly called virtual or remote live proctoring but can also encompass assisted online proctoring. For the sake of clarity and consistency, this paper utilizes the terms established by the National College Testing Association (NCTA). As defined by NCTA (2022), in-person proctoring involves a proctor supervising a test event in real time at the same location as the test-taker. Online proctoring is done in real time, with a remote proctor verifying the identity of the candidate and observing the candidate throughout the exam using various technologies. Assisted online proctoring is the same as online proctoring but also employs artificial intelligence (AI) or algorithms to alert the online proctor of possible irregularities occurring in real time during the testing session. Monitoring involves video recording of the examination with the option to have a proctor or other individual review the recording after the fact to identify any testing irregularities. In addition to recording the testing session, assisted monitoring also employs AI or algorithms to identify irregularities. These concerns are flagged in the system as possible fraud or cheating for later review by a human proctor (NCTA, 2022).

The research into the different proctoring modes, especially in higher education, has primarily focused on reviews of the different types of proctoring and monitoring systems (Cramp et al., 2019; Draaijer, 2017; Fiano et al., 2021; Hussein et al., 2020; Nigam et al., 2021; Raman et al., 2021), security (Langenfeld, 2020), privacy (Coghlan et al., 2021), fairness (Langenfeld, 2020), testing environments (Traylor et al., 2021) and/or user experiences (Fatima et al., 2021; Halem et al., 2021; Kharbat & Daabes, 2021; Kolski & Weible, 2018; Lilley et al., 2016; Raman et al., 2021; Woldeab & Brothen, 2021). Studies exploring the impact of different proctoring modes on test performance are not as common, and those that do exist tend to focus on proctored (in-person or online) versus unproctored exams. Many of these studies reported no differences between proctored and unproctored exam scores (Hollister & Berenson, 2009; Jaap et al., 2021; Karim et al., 2014; Kharbat & Daabes, 2021; Ladyshewsky, 2015; Rios & Liu, 2017). These studies cautiously concluded that remote delivery of testing without proctoring may be an acceptable option for some examinations. In contrast, other studies revealed that mean or median scores were significantly higher for unproctored exams compared to proctored exams (Alessio et al., 2017; Daffin Jr. & Jones, 2018; Goedl & Malla, 2020; Reisenwitz, 2019; Steger et al., 2020). Since cheating was one of the likely causes for the differences, the general recommendation was that most exams be proctored, especially high stakes exams; however, it was acknowledged that unproctored exams may be feasible for certain tasks (Steger et al., 2020).

Studies that compare specific proctoring modes and test performance are scarce but germinating due to the exponential growth in online proctoring and monitoring during the pandemic. One such study conducted before the pandemic by Weiner and Hurtz (2017) compared scores on professional licensing exams proctored online or in-person at test centers during a one-year time frame. This study included data from three exams completed by 14,623 candidates, of which 6,889 were proctored online and 7,734 were proctored in-person at test centers. Candidates who were proctored online completed their exams at kiosk computers located in college libraries, testing labs, office buildings, and community centers, usually in an enclosed room. The results revealed no significant differences between proctoring modes on all three exams. The explanation provided by Weiner and Hurtz (2017) for the nominal impact on test scores was that the kiosk-based testing environment closely mimicked that of an in-person testing environment.

A study by Prigoff and colleagues (2021) conducted during the pandemic involved comparing historical test scores (pre-pandemic) proctored in-person to test scores proctored online during the pandemic. In this study, several performance measures were evaluated, but only one, the National Board of Medical Examiners (NBME) shelf exam, was proctored online as compared to in-person. Comparing the historical performance of students proctored in-person over three years prior to the pandemic (n = 61) to that of students proctored online in 2020 (n = 19) revealed a higher mean score on the NBME shelf exam for the online proctored group compared to the in-person proctored group, but the difference was not significant (Prigoff et al., 2021). A description of the online proctoring was not provided in this study, nor was a description of the NBME shelf exam. This, plus the small sample size at a single institution limits the interpretation and generalization of these findings.

Another study conducted during the pandemic by Cherry and colleagues (2021) investigated the equivalency of professional licensure examinations administered via online proctors and in-person proctors at certified test centers across four American states from May to December of 2020. The 11 exams were all multiple-choice format and varied from 90-150 items per exam. Sample sizes varied from 746-1,863 candidates per exam, with the total number of candidates in excess of 14,000. Across proctoring modes, the online proctored groups were smaller (n = 213-831) compared to the in-person proctored groups (n = 484-1,070) for all but one exam. The online proctoring appears to have been provided by ProProctor, an online proctoring service offered by Prometric. Although it is not specified in the paper, according to the ProProctor website, candidates are monitored continuously with at most an 8:1 candidate to proctor ratio (Prometric, 2022). Although some minor differences were detected for specific exams, overall, Cherry and colleagues (2021) found no significant difference between the outcomes for candidates across the two proctoring modalities. The authors acknowledged that the findings were specific to insurance industry exams but still concluded that their research provided "an early indication that outcomes from tests administered…in testing centers and remotely, via LRP [Live Remote Proctors], are equivalent" (Cherry et al., 2021, p.8).

A study comparing proficiency test scores by proctor type for applicants applying to a medical program by Andreou and colleagues (2021) revealed similar

results. A total of 593 candidates completed the admission exam, with 472 proctored online and 121 proctored in-person. In this study, assisted online proctoring was employed so any suspicious behavior flagged by the AI technology could be investigated immediately by the proctor. Once again, test scores did not differ significantly, so the researchers concluded that "proctoring type does not influence exam results" (Andreou et al., 2021, p.7).

Contrasting results were reported by Wuthisatian (2020) in a study that compared final exam scores of students enrolled in an online MBA program. The final exam was administered either onsite at an accredited educational institution, authorized testing center, or public library, or proctored online using a professional testing service (e.g., ProctorU). For the latter, the students were allowed to take the test anywhere on their personal computer, but their computer had to have the required technology so the online proctor could supervise them throughout the entire testing session. Of the 65 students who took the exam, 36 tested with the online proctor, and 29 tested in-person. Regression analysis revealed a significant difference in average scores, with students proctored in-person having a significantly higher average test score compared to students proctored online. Further analysis revealed that students' unfamiliarity with online proctoring procedures and requirements may have factored into the lower scores for the online proctored group. To explain this, Wuthisatian (2020) postulated that students' inexperience with online proctoring may have increased their test anxiety and stress levels, especially if they experienced technical difficulties during the testing session, which likely weakened their test performance.

Findings from these initial studies are encouraging as, overall, they seem to demonstrate that online proctoring is a viable alternative to in-person proctoring. However, substantially more studies are required before any finite conclusion can be determined. Testing is a value-laden activity with scores being used to make consequential decisions, so it is vital that test scores are not affected by proctor modality. For example, in higher education there is a need to conduct more research related to admission and placement testing, because even minor differences in test scores by proctor modality could result in errors in admission and/or placement into post-secondary institutions. To avoid this, all online proctoring services utilized by post-secondary institutions, commercial or in-house, need to be evaluated to ensure consistency of procedures and outcomes. Research that explores if and how test format (e.g., static, computer adaptive), item format (e.g., multiple choice, open answered), subject matter, purpose (e.g., certification, ability, knowledge), and time constraint factor into performance by proctor modality also needs to be conducted. Studies exploring changes in the students' level of comfort and experience with online proctoring post-pandemic could prove informative too. Although it is impossible to investigate all of these factors at once, studies can start filling in the gaps by focusing on at least some. This study is equipped to do that by comparing assessment scores for placing and admitting students in higher education by proctor modality.

METHOD

This study was conducted at Thompson Rivers University (TRU), a public, open access Canadian university that offers undergraduate and master's degrees as well as certificates and diplomas (TRU, 2021). As an open access university, admission for most students is based on a first come first serve basis with minimal requirements. The most common requirement is English language proficiency, which can be demonstrated through coursework (high school or upgrading) or English assessment test scores. Some programs also require applicants to demonstrate mathematics proficiency, again through coursework or mathematics assessment test scores. The TRU Assessment Centre facilitates the English and mathematics entry assessment testing utilizing the ACCUPLACER® testing platform.

Entrance Assessment Tool - ACCUPLACER®

ACCUPLACER® is a web-based testing platform that offers a series of tests to assess students' academic skills in English and mathematics. Approximately 1,200 post-secondary institutions in North America utilize ACCUPLACER® tests for advising, placement, and/or admission purposes (K. Montognese, personal communication, April 11, 2022). Most of the tests are multiple-choice except for the WritePlacer® and WritePlacer ESL® tests, which require candidates to complete an on-demand essay to a specific prompt. The multiple-choice exams offered by this online testing system are computer-adaptive, so the questions and the sequence in which they are presented vary from student to student. Specifically, the first question of each multiple-choice test is moderately difficult and randomly selected. If the student answers correctly, the next question or set of questions presented will be slightly more difficult, while an incorrect answer results in an easier question or set of questions being presented (College Board, 2018a). This pattern repeats itself throughout the exams, making each test capable of assessing a wider range of student abilities in a reduced testing time with immediate feedback and flexible testing sessions. The WritePlacer® test randomly assigns one of over a dozen prompts provided by the ACCUPLACER® system. Students submit a response, which is scored utilizing automated assessment technology (College Board, 2018b). The ACCUPLACER® tests can be administered in-person at qualified testing centers or online through College Board-approved online proctoring vendors and other testing professionals.

At TRU, six tests are utilized to assess the English and mathematics skills of applicants: Reading; Writing; WritePlacer®; Arithmetic; Quantitative Reasoning, Algebra, and Statistics (QAS); and Advanced Algebra and Functions (AAF; College Board, 2017). The scoring for the five multiple-choice tests ranges from 200-300, while the scoring for the WritePlacer® test ranges from 0-8 (College Board, 2017). At TRU, all tests are untimed except for the WritePlacer®, which has a one-hour time limit. Applicants may take just one test or some combination of all six tests depending on what they require for entry into TRU.

ACCUPLACER® Testing at TRU during the COVID-19 Pandemic

Prior to the COVID-19 pandemic, ACCUPLACER® tests were administered in-person either at the TRU Assessment Centre or another approved external testing site (e.g., other university or college testing

center, high school). At the beginning of the pandemic, the TRU Assessment Centre closed briefly for the month of April 2020 but resumed in-person testing activities in May of 2020 under strict health and safety protocols. These protocols included social distancing, mask mandates, hand sanitizing, and limits on testing volume. The external testing sites also closed, and most remained closed to external in-person testing for much longer than TRU. With these external testing sites unavailable and in-person testing at TRU limited by the pandemic protocols, the Assessment Centre turned to one of the online proctoring services approved by the College Board, Examity®, in order to meet the demand for ACCUPLACER® testing. This vendor provides online proctoring services year-round, 24 hours per day by trained proctors (Examity, 2022a). Online proctoring by Examity® for ACCUPLACER® testing involves a proctor supervising a maximum of two students per session (Examity, 2022b; K. Montognese, personal communication, April 11, 2022).

The TRU Assessment Centre is a small testing center with four staff members, all of whom have been involved in high stakes testing for a minimum of ten years. The Centre is a Certified Test Center by the NCTA, initially certified in 2016 and recertified in 2021. This certification indicates that the TRU Assessment Centre demonstrates mastery of best practices in the testing industry, operating and adhering to the professional standards for all types of testing (NCTA, 2021).

Participants

Data for this study were collected from 918 participants who completed an English and/or math assessment for TRU from May 2020 to February 2022. This time frame was utilized to ensure a consistent testing environment for in-person testing. From reopening in May 2020 until February 2022, the Assessment Centre's pandemic protocols remained in place. In mid-March 2022, the health and safety mandates were removed, so students no longer had to wear masks, socially distance, or sanitize before testing, although they were still encouraged to do so. To control for other confounding variables, there was no overlap of students between proctored groups. Within the groups, only one session of testing was included in the study, so all rewrites were excluded. Of the 918 participants who met these requirements, 561 participants were proctored in-person by TRU staff, and 357 were proctored online by Examity® employees. The average age of the in-person proctored group was 24.5 (SD = 8.34) compared to 26.1 (SD = 9.76) for the online proctored group. For the in-person proctored group, 35% identified as female, 63% identified as male, and 2% choose not to answer. For the online proctored group, 35% identified as female, 41% identified as male and 24% chose not to answer.

Data Analysis

To explore the pattern of testing by proctor type, in-person versus online, tallies for both were calculated by month and compared visually in a line graph. Descriptive statistics for all tests by proctor group were tabulated. The distribution of test scores for both groups were assessed using the Shapiro-Wilk Test of Normality. Based on this test, the Writing scores for the group that was proctored in-person were normally distributed, but all other onsite testing scores were not. For the online proctored groups, the Reading and QAS test scores were normally distributed, but scores for the other four tests were not. Hence, the Mann-Whitney test was employed to compare test scores between the in-person

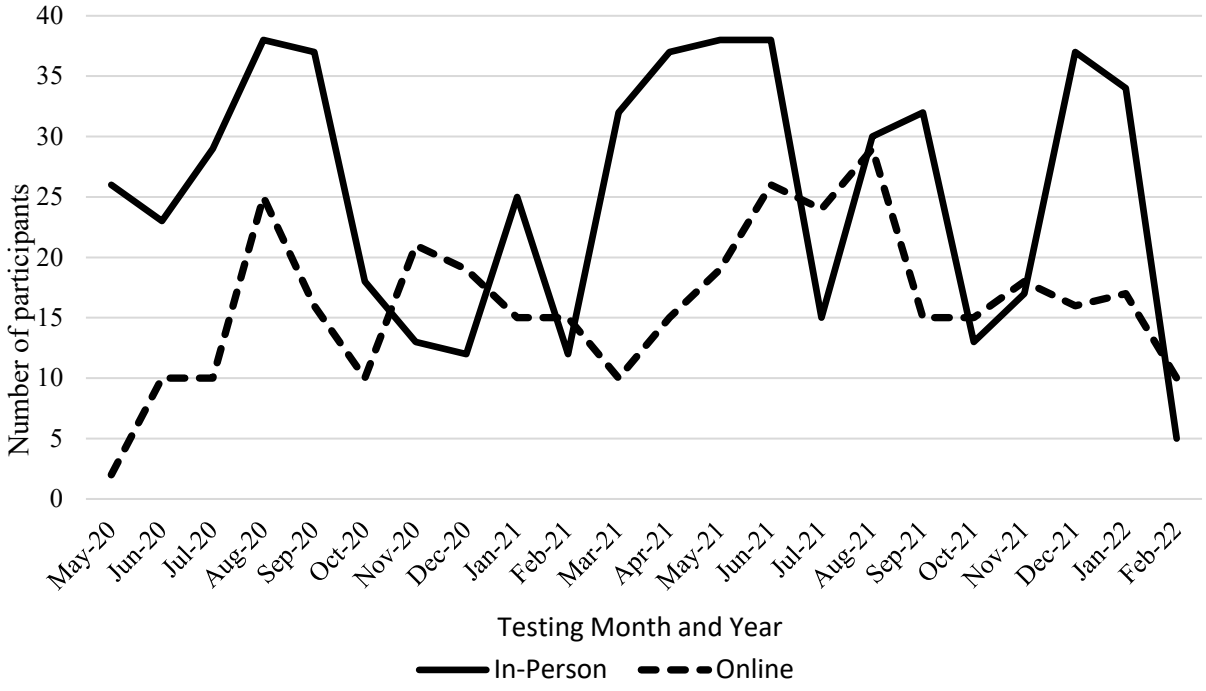
and the online proctored groups. No assumption was made about differences between test scores; therefore, two-tailed tests were conducted with the level of significance set to .05. The effect size for the differences between test scores also was calculated (ES = $|z| / \sqrt{N}$).

RESULTS

Testing at TRU during the pandemic by type of proctoring varied greatly (Figure 1). At the beginning of testing during the pandemic (May 2020), online proctoring was just being established, so the volume of testing for this group was limited. After that, it fluctuated much like in-person testing.

Figure 1

Monthly Testing Volume by Proctor Mode



The distributions of the ACCUPLACER® English tests by proctor type were very similar (Table 1). The mean score for the ACCUPLACER® Reading test was slightly greater for the in-person proctored group as compared to the online group, but the medians were identical. The same held true for the ACCUPLACER® Writing mean and median test scores, with the in-person proctored group scoring marginally higher than the online group. For the ACCUPLACER® WritePlacer® essay, mean scores were almost the same across the two groups, but the median score was different by one point on an eight-point scale in favor of the online proctored group.

Table 1

Descriptive ACCUPLACER□ English Test Scores by Proctoring Group

	Reading		Writing		WritePlacer□	
Statistics	In-Person	Online	In-Person	Online	In-Person	Online
Mean	260.75	260.22	257.07	256.87	4.51	4.57
Median	259.50	259.50	257.00	256.00	4.00	5.00
SD	16.692	15.155	12.190	13.918	1.067	.966
Maximum	300	300	295	300	7	7
Minimum	200	219	211	200	1	2
n	462	228	270	155	107	109

Table 2

Descriptive ACCUPLACER□ Mathematics Test Scores by Proctored Group

	Arithmetic		Quantitative Reasoning, Algebra & Statistics (QAS)		Advanced Algebra & Functions (AAF)	
Statistics	In-Person	Online	In-Person	Online	In-Person	Online
Mean	271.27	269.11	256.42	259.73	240.94	249.96
Median	270.00	268.00	259.00	260.00	243.00	254.40
SD	15.141	16.411	16.756	17.765	25.570	23.752
Maximum	300	300	300	300	300	300
Minimum	223	213	211	211	200	200
n	476	256	438	237	111	134

The distributions of the ACCUPLACER® mathematics tests varied more so than the English test scores (Table 2). The mean score for the in-person proctored group was slightly higher, as was the median, compared to the online proctored group for the ACCUPLACER® Arithmetic test. The opposite was true for the ACCUPLACER® QAS exam, as the online proctored group's mean and median

scores were slightly higher than that of the in-person proctored group. The ACCUPLACER® AAF test exhibited the most variance of all, with mean and median scores for the online proctored group being greater than the in-person proctored group (Table 2).

Based on Mann Whitney tests, there were no significant differences in the test scores for the group of students who were proctored in-person as compared to the group of students who were proctored online, except for the AAF test (Table 3). For the AAF, there were significant differences between the scores, with the participants who were proctored online scoring higher than those who were proctored in-person at the TRU Assessment Centre ($z = 2.839$, $p = .005$). The effect size of the difference was small (ES = 0.18; Table 3).

Table 3

Comparison of Test Scores by Proctored Group: Mann-Whitney Test Results

ACCUPLACER□ test	In-Person		Online		Test Statistics		
	n	Mean Rank	*n*	Mean Rank	*z*	*p**	*Effect Size*
Reading	462	347.53	228	341.53	-0.381	.703	.0145
Writing	270	213.93	155	211.44	-0.199	.843	.0010
WritePlacer□	107	107.44	109	109.54	0.261	.794	.0176
Arithmetic	476	375.97	256	348.89	-1.653	.098	.0611
QAS	438	327.84	237	356.78	1.841	.066	.0709
AAF	111	108.89	134	134.69	2.839	**.005**	**.1814**

*two-tailed.

DISCUSSION

Sudden reliance on online proctoring services due to the COVID-19 pandemic was and still is a concern for educators, especially testing professionals in higher education. The primary concerns are related to privacy, academic integrity, access, and performance. This study focused on the latter, comparing the performance of students proctored in-person at a testing center to that of students proctored online by a remote agent. Results indicated performance was not significantly impacted by proctoring modality for all three English tests and two of the three mathematics tests. For the third mathematics test, Advanced Algebra and Functions, there was a significant difference, but the effect size was small.

This study reaffirms the findings from previous studies (Andreou et al., 2021; Cherry et al., 2021; Prigoff et al., 2021; Weiner & Hurtz, 2017) that proctoring mode does not impact test performance, at

least not during the pandemic. The similarity between proctoring modes provides the best explanation for these results. Online proctoring closely matches in-person proctoring in terms of roles and procedures, with the most important similarity being the continuous, real-time proctoring of each session. The low candidate to proctor ratio (average of 2:1) offered by Examity® may also factor into these results as it probably facilitates better support for candidates while protecting the integrity of the exam (Examity, 2022c).

Testing during the pandemic may also provide some insights into these results. Prior to the pandemic, students were more familiar and likely more comfortable with in-person proctoring than online proctoring. However, the switch to virtual learning environments during the pandemic and introduction of safety protocols such as wearing a mask, sanitizing, and social distancing for in-person testing likely reversed that trend. Testing in-person became just as stressful, maybe even more so, than testing online in the security of one's own home. Also, during the pandemic students became immersed in virtual learning and thus more accustomed to the technology utilized for online proctoring of exams.

The testing tool also may have contributed to these findings. The ACCPULACER® testing platform is computer adaptive, so, as noted in the Methods section, questions on the multiple-choice tests are selected at random and depend upon the answer to the previous question. Hence, it is very difficult to predict the next question, thereby limiting cheating opportunities as compared to a static multiple-choice test. The same may be true for the essay-type questions, as they did not differ significantly by proctoring mode. Once again, the ACCUPLACER® testing platform offers different essay prompts that are randomly assigned to candidates, which limits cheating more so than static, single prompt essay tests. In addition, the ACCUPLACER® system includes a lockdown browser to increase test security. If a student clicks outside the Test Administration window while their session is in progress, the student is automatically locked out of the test until the proctor reauthorizes access to the test (College Board, 2018b). These features improve the efficacy of proctoring in-person and online, allowing for a similar testing experience regardless of proctoring mode.

Explaining the findings related to subject matter is more problematic. In this study, the results demonstrated that there were no differences between English test scores for the in-person and online proctored groups. However, for the mathematics tests there was a significant difference, with the online proctored group scoring significantly higher on the AAF test. Although not significant but notable, the online proctored group also scored higher on the QAS. Meanwhile, the opposite was true for the Arithmetic test. Hence, it is not clear if subject matter has an impact on test score differences by proctor type. In terms of the mathematics tests, perhaps the difficulty of the test content played a role, since the AAF and QAS tests cover more difficult concepts than the Arithmetic test (College Board, 2017). Whatever the reason, more research into subject differences is needed to verify and explain the findings.

Limitations and Future Research

There are several limitations that hinder potential generalizations and implications of this study. First, the data is specific to one institution—a public, open access university in Canada—so the findings may not apply to other post-secondary institutions with

different populations, programs and/or admission processes. Second, the data only applies to test scores from the ACCUPLACER® testing system, so it may not be applicable to other assessment tools utilized by other institutions. Hence, this study needs to be replicated at other post-secondary institutions, especially those that utilize different assessment tools.

Third, this study focused on online proctoring offered by a private vendor. Consequently, the findings may not apply to the other modes of online proctoring and monitoring (e.g., assisted online, monitoring, and assisted monitoring) offered by different vendors, nor online proctoring being conducted by educational testing staff and faculty using a myriad of different platforms. The efficacy of these different variations of online proctoring and monitoring needs to be examined in future studies.

Finally, it is also critical to repeat this study once the pandemic is over, and higher education testing services return to "normal," as this will result in more changes that need to be investigated. For example, will the removal of the pandemic protocols such as masking, social distancing, or sanitizing alter performance for in-person testing sessions?

Conclusion

As the worldwide COVID-19 pandemic swept through higher education, online proctoring of exams surged, supplementing and in some cases replacing traditional in-person proctoring. The assumption is that online proctoring equates to the gold standard of in-person proctoring. Research needs to verify this assumption, especially in terms of outcomes. This study addresses the dearth of existing research by comparing assessment test scores from in-person proctored testing to online proctored testing. Based on these findings, it appears as though the performance of students on tests proctored in-person or online, on the whole, do not vary substantially. While these findings are promising, they need to be confirmed or refuted by subsequent studies.

REFERENCES

Alessio, H. M., Malay, N., Maurer, K., Bailer, A. J., & Rubin, B. (2017). Examining the effect of proctoring on online test scores. *Online Learning, 21*(1), 146–161.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andreou, V., Peters, S., Eggermont, J., Wens, J., & Schoenmakers, B. (2021). Remote versus on-site proctored exam: Comparing student results in a cross-sectional study. *BMC Medical Education, 21*(1), 624. https://doi.org/10.1186/s12909-021-03068-x

Cherry, G., O'Leary, M., Naumenko, O., Kuan, L.-A., & Waters, L. (2021). Do outcomes from high stakes examinations taken in test centres and via live remote proctoring differ? *Computers and Education Open, 2*(100061-). https://doi.org/10.1016/j.caeo.2021.100061

Coghlan, S., Miller, T., & Paterson, J. (2021). Good proctor or "big brother"? Ethics of online exam supervision technologies. *Philosophy & Technology, 34*(4), 1581–1606. https://doi.org/10.1007/s13347-021-00476-1

College Board. (2017). *Next-generation test specifications: Version 2.0*. https://accuplacer.collegeboard.org/accuplacer/pdf/next-generation-test-specifications-manual.pdf

College Board. (2018a). *ACCUPLACER program manual*. https://accuplacer.collegeboard.org/accuplacer/pdf/accuplacer-program-manual.pdf

College Board. (2018b). *ACCUPLACER user's guide*. https://accuplacer.collegeboard.org/accuplacer/pdf/accuplacer-user-guide.pdf

Cramp, J., Medlin, J. F., Lake, P., & Sharp, C. (2019). Lessons learned from implementing remotely invigilated online exams. *Journal of University Teaching and Learning Practice, 16*(1).

Daffin Jr., L. W., & Jones, A. A. (2018). Comparing student performance on proctored and non-proctored exams in online psychology courses. *Online Learning, 22*(1), 131–145.

Draaijer, S. (2017). *START REPORT: A report on the current state of online proctoring practices in higher education within the EU and an outlook for OP4RE activities*. Erasmus+ OP4RE project.

Examity. (2022a). *Administer ACCUPLACER with Examity*. https://www.examity.com/features/accuplacer/

Examity. (2022b). *History*. https://www.examity.com/about/history/

Examity. (2022c, June). *Creating a positive test-taker journey*. Https://www.examity.com/Creating-a-Positive-Test-Taker-Journey/

Fatima, S. S., Idrees, R., Jabeen, K., Sabzwari, S., & Khan, S. (2021). Online assessment in undergraduate medical education: Challenges and solutions from a LMIC university. *Pakistan Journal of Medical Sciences, 37*(4), 945–951. https://doi.org/10.12669/pjms.37.4.3948

Fiano, K. S., Medina, M. S., & Whalen, K. (2021). The need for new guidelines and training for remote/online testing and proctoring. *American Journal of Pharmaceutical Education, 85*(8), 805–808.

Goedl, P. A., & Malla, G. B. (2020). A study of grade equivalency between proctored and unproctored exams in distance education. *American Journal of Distance Education, 34*(4), 280–289.

Halem, N., Klaveren, C., & Cornelisz, I. (2021). The effects of implementation barriers in virtually proctored examination: A randomised field experiment in Dutch higher education. *Higher Education Quarterly, 75*(2), 333–347.

Hollister, K. K., & Berenson, M. L. (2009). Proctored versus unproctored online exams: Studying the impact of exam environment on student performance. *Decision Sciences Journal of Innovative Education, 7*(1), 271–294. https://doi.org/10.1111/j.1540-4609.2008.00220.x

Hussein, M. J., Yusuf, J., Deb, A. S., Fong, L., & Naidu, S. (2020). An evaluation of online proctoring tools. *Open Praxis, 12*(4), 509–525. https://doi.org/10.5944/openpraxis.12.4.1113

Jaap, A., Dewar, A., Duncan, C., Fairhurst, K., Hope, D., & Kluth, D. (2021). Effect of remote online exam delivery on student experience and performance in applied knowledge tests. *BMC Medical Education, 21*(1), 86. https://doi.org/10.1186/s12909-021-02521-1

Karim, M. N., Kaminsky, S. E., & Behrend, T. S. (2014). Cheating, reactions, and performance in remotely proctored testing: An exploratory experimental study. *Journal of Business and Psychology, 29*(4), 555–572.

Kharbat, F. F., & Daabes, A. S. A. (2021). E-proctored exams during the COVID-19 pandemic: A close understanding. Education and Information Technologies, 26(6), 6589–6605.

Kimmons, R., & Veletsianos, G. (2021). Proctoring software in higher ed: Prevalence and patterns. In *EDUCAUSE Review*. https://er.educause.edu/

Kolski, T., & Weible, J. (2018). Examining the relationship between student test anxiety and webcam based exam proctoring. *Online Journal of Distance Learning Administration, 21*(3).

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. https://doi.org/10.1080/02602938.2014.956683

Langenfeld, T. (2020). Internet-based proctored assessment: Security and fairness issues. *Educational Measurement: Issues and Practice, 39*(3), 24-27. https://doi.org/10.1111/emip.12359

Lilley, M., Meere, J., & Barker, T. (2016). Remote live invigilation: A pilot study. *Journal of Interactive Media in Education, 2016*(1), 1-5.

MeazureLearning. (2022). *Higher education* (Vol. 2022, Issue April 10,). https://www.meazurelearning.com/markets/higher-education

National College Testing Association. (2021). *Test center certification*. https://www.ncta-testing.org/test-center-certification

National College Testing Association. (2022). *Proctoring and monitoring definitions*.

Nigam, A., Pasricha, R., Singh, T., & Churi, P. (2021). A systematic review on AI-based proctoring systems: Past, present and future. *Education & Information Technologies, 26*(5), 6421–6445. https://doi.org/10.1007/s10639-021-10597-x

Prigoff, J., Hunter, M., & Nowygrod, R. (2021). Medical student assessment in the time of COVID-19. *Journal of Surgical Education, 78*(2), 370–374. https://doi.org/10.1016/j.jsurg.2020.07.040

Proctorio. (2022). *Proctorio history*. https://proctorio.com/about/history

Prometric. (2022). *ProProctor*. https://www.prometric.com/proproctor

Raman, R., B, S., G, V., Vachharajani, H., & Nedungadi, P. (2021). Adoption of online proctored examinations by university students during COVID-19: Innovation diffusion study. *Education & Information Technologies, 26*(6), 7339–7358. https://doi.org/10.1007/s10639-021-10581-5

Reisenwitz, T. H. (2019). Examining the necessity of proctoring online exams. *Society for Marketing Advances Proceedings*, 515–516.

Rios, J. A., & Liu, O. L. (2017). Online proctored versus unproctored low-stakes internet test administration: Is there differential test-taking behavior and performance? *American Journal of Distance Education, 31*(4), 226–241.

Steger, D., Schroeders, U., & Gnambs, T. (2020). A meta-analysis of test scores in proctored and unproctored ability assessments. European Journal of Psychological Assessment, 36(1), 174–184. https://doi.org/10.1027/1015-5759/a000494

Thompson Rivers University. (2021). *TRU Institutional Accountability Plan and Report 2020-2021*.

Traylor, Z., Hagen, E., Williams, A., & Arthur, W. (2021). The testing environment as an explanation for unproctored internet-based testing device-type effects. *International Journal of Selection & Assessment, 29*(1), 65–80. https://doi.org/10.1111/ijsa.12315

Weiner, J., & Hurtz, G. (2017). A comparative study of online remote proctored versus onsite proctored high-stakes exams. *Journal of Applied Testing Technology, 18*(1), 13–20. http://www.jattjournal.com/index.php/atp/article/view/113061

Woldeab, D., & Brothen, T. (2021). Video surveillance of online exam proctoring: Exam anxiety and student performance. *International Journal of E-Learning & Distance Education, 36*(1).

Wuthisatian, R. (2020). Student exam performance in different proctored environments: Evidence from an online economics course. *International Review of Economics Education, 35*. https://doi.org/10.1016/j.iree.2020.100196