ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #49535

Stepping Back from a Digital Age: Paper and Pen Coding Exams in a post GenAI world

Mr. Lance Leon Allen White, Texas A&M University

Lance White is a Ph.D. student at Texas A&M University in Interdisciplinary Engineering with a thrust in Engineering Education. He is working as Lecturer for the Engineering Academic and Student Affairs group in the College of Engineering.

Stepping Back from a Digital Age: Paper and Pen Coding Exams in a post GenAI world

Introduction

Generative AI (GenAI) has fundamentally altered the educational landscape, bringing both advantages and challenges. In engineering education, the rapid adoption of GenAI tools has facilitated learning but has also spurred a notable increase in academic dishonesty. In the wake of this shift researchers have been quick to examine effects. Chan [1] explored this phenomena and introduced the concept of "AI-giarism", describing the misuse of AI tools to bypass traditional plagiarism detection systems through a qualitative study of over 500 students. Li [2] emphasizes in their work the growing ethical dilemmas stemming from hard to monitor usage of GenAI in assessments, ultimately calling for adaptive educational policies to address this issue. It is clear that higher education is aware that this is a significant problem, and work by the authors explored this early in this shift [3-5].

Academic dishonesty in the context of STEM disciplines has seen a notable rise. To set a baseline understanding we have to refer to earlier works such as McCabe and Trevino [6] identifying contextual influences like institutional policies an honor codes as significant factors of shaping student behavior. Texas A&M University (TAMU) has within it a dedicated unit to handle these issues, and the university motto is significantly tied to an identity of pride and honorable behavior. In more recent work Newton and Essex [7] notes a significant uptick in reports of AI-assisted misconduct in higher education, particularly in STEM disciplines, where problem-solving tasks are more susceptible to automation by AI tools. Literature asserts that currently GenAI's capabilities have complicated efforts to uphold academic integrity.

The work by Chan [1] and Li [2] serve to demonstrate GenAI's ability to produce sophisticated, human-like responses making traditional plagiarism detection tools less effective. Such trends underscore the urgency for innovative strategies to safeguard the integrity of educational assessments and ensure equitable evaluation practices. Previous work by the authors examined faculty perceptions of the likelihood of academic misconduct pre- and post-ChatGPT's release and found that faculty overall were very hesitant and skeptical of the technology early in its release, with sentiments changing one year after the original survey [3-5]This paper addresses the challenges associated with academic dishonesty and GenAI within the context of a widely offered first-year engineering course.

TAMU's first-year engineering program has been facing the academic dishonesty challenge directly. Exams for this course since the beginning of the COVID-19 pandemic were transitioned completely to digital exams via the Canvas learning management system. During the pandemic these exams were proctored via Zoom. When students eventually returned to the classroom, these exams were proctored in-person in a Canvas exam session. This option offered convenience in grading and enabled simple and irrefutable methods of detecting academic

misconduct. With the rise of tools like ChatGPT, faculty noticed a perceptible increase in unethical practices resulting in academic dishonesty proliferation throughout the first-year engineering student population. This necessitated a re-evaluation of assessment methodologies.

The first-year engineering cohort of faculty transitioned from the digital to paper format during Summer 2024 in a small scale, resulting in no academically dishonest behaviors in that small population. This success positioned the faculty to employ this method of assessment into the standard procedures of the academic unit.

It is important to state the context in which this course of interest is situated. There are a total of three courses required by the college of engineering at TAMU prior to admittance to an engineering major. Students are rewarded for high performance (3.75 GPA at the time of major application) with an auto-admission into their first choice of a major through this process. The possibility of automatic acceptance is a major driver in these students' lives as they navigate this first year. While there is no quantitative or qualitative work looking at this phenomenon at TAMU. The authors are aware of this through discussions with their students, administration, and staff in the college of engineering as a cultural norm. The authors of this work see this as a potential driving motivation for students to engage in academically dishonest activities at any rate beyond zero.

Literature

Academic dishonesty has been a persistent challenge in higher education, with digital assessment methods amplifying its visibility and complexity. Several studies provide valuable insights into the comparative dynamics of digital versus paper-based exams. McCabe and Trevino [6] provide foundational insights into how institutional culture can influence dishonesty, which in the case of TAMU, at least on the surface, should be a culture of zero-tolerance of the behavior. Gallant [8] further explored the role of teaching praxis an institutional policy, finding that these components played significant roles in creating a culture that prioritizes ethical behavior. Research by Holden, et al. [9] shows that digital exams often result in higher instances of cheating compared to paper-based assessments, attributed to the availability of online resources and the difficulty of effective proctoring. Work by Dendir and Maxwell [10] highlight students’ perceptions of paper exams being more secure and less prone to dishonest practices, fostering a stronger sense of academic integrity. Some of the more recent works in this space by Chan [1] and Li [2] discuss the role of GenAI tools in bypassing traditional plagiarism detection, creating an urgent need for adaptive assessment methodologies rather than the accepted standards of the past.

Work by Dendir and Maxwell [10] does set a stage similar to which our work is positioned wherein they found student performance to not differ between digital and paper formats, rather the student attitudes and engagement levels were significantly impacted by assessment method. Stowell and Bennett [11] conducted a study that did indicate that online testing does increase test anxiety and thus induces a situation in which misconduct is deemed appropriate in the minds of

students enduring through this anxiety. We do see a gap in direct comparisons in a post-GenAI world, especially in the context of first-year engineering courses where stakes are somewhat hirer for students in comparison to some other higher-education programs.

This body of literature provides a comprehensive foundation for examining the transition from digital to paper-based exams, particularly in the context of addressing academic dishonesty in higher education. By integrating insights from various studies, it highlights the comparative dynamics of assessment formats and their implications for both student behavior and institutional practices. This context enriches the analysis of the methods and outcomes observed in the first-year engineering program at TAMU, situating the findings within a broader academic discourse and paving the way for future explorations in this evolving field.

Methods

The authors of this paper have taken the threat of GenAI enabled academic dishonesty seriously and have migrated the entirety of our first-year engineering courses from digital to paper exams. This change in assessment was piloted during the Summer of 2024 and extended to the whole of the first-year engineering courses in the Fall of 2024. The exams between years are very similar, both being developed by one of the authors of this work with input from the faculty. The grades examined in this study are from the authors' course where populations for each year are independent of one another. There is argument that some differences might be explained by improved teaching of the authors, but for this study this will not be accounted for. The student demographics are largely the same, and representative of the typical distribution of student identities and backgrounds for the college of engineering at TAMU. The mode of assessment was not foreign to the students prior to attempting the exams as quizzes were conducted via paper to prepare students for that mode of assessment both as a practice for the content of the exams and the format of the exams. A total of 560 and 590 students completed Exam 1 for the Fall 2023 and Fall 2024 semesters respectively, while 559 and 583 students completed Exam 2. The difference here is explained either by students dropping the class, failing to attend the exam, or withdrawing from the university.

Academic dishonesty data is significantly difficult to obtain through the office responsible for handling this behavior. Considering this lack of transparency available through TAMU's systems and offices, the authors compare their own rates of incidence and reported rates of incidence from colleagues also teaching this course. While these reporting numbers are far from large enough to conduct any statistical analysis, the rates will be discussed in an experiential format, pulling from the lived experiences of the authors.

Initial inspection of the data revealed some potential inconsistencies as zero values exist in the final reporting of exam grades for students who failed to appear and complete some exams. These values were removed from the data to maintain the integrity of the analysis following best practice (Osborne & Overbay, 2004). Two modes of analysis were used for this study,

descriptive statistics and visualization along with independent t-tests. Summary statistics are presented as an overview of score distributions. Histograms and box and whisker plots are used to succinctly present the data to visually identify the changes in performance between groups (Tukey, 1977). Independent t-tests were necessary due to the student populations being distinctly different and the sample sizes being unmatched. Two independent two-sample t-tests were used to assess the mean differences in scores between the two years and determine significance. The null hypothesis in these cases was that no significant difference would exist in mean scores from year to year, substantiated by the control of faculty and assessment design. A 95% confidence level is used meaning that a p value less than 0.05 would indicate significant difference in the mean.

The framework used to perform this analysis draws on established data analytical methods in education. Cleaning the data aligns with recommendations by Osborne & Overbay (2004). The choice to use independent t-tests is widely considered standard in educational studies to compare group means as well (Cohen, Manion, & Morrison, 2011).

This analysis was conducted with Python using the pandas and scipy libraries while visualizations call on the matplotlib library. These tools are largely recognized for their reliable codebases and transparency for analysis (McKinney, 2010). Canvas was the learning management system used to host the exams for 2023 while Gradescope was used to analyze the paper exam submissions for 2024, scanned in and submitted by the authors.

Results

Exam Comparison

Using the methods discussed above student performance from 2023 to 2024 improved between both exams. These changes can be seen below when comparing the distributions and box and whisker plots of the two exams across the two years. In Figure 1 the distributions from 2023 and 2024 are compared, showing relatively similar distribution shapes while in Figure 2 we can see the median and quartile 2 to 3 to be very similar. The mean for these two semesters Using a two-sample t-test the comparison between Fall 2023 and Fall 2024 to have a p-value of 0.05, suggesting weak evidence to reject the null hypothesis.

Exam 2 when compared provides a markedly different story. Figure 3 showcases the distributions with significantly different shapes, with Fall 2024 students well outperforming the Fall 2023 cohort. This is seen more succinctly in Figure 4 where the median values are vastly different.

Table 1 presents the descriptive statistics for each of the exams, further suggesting a significant increase for Exam 2. A two-sample t-test to compare Exam 2 results in a p-value of 6.9E-16, suggesting a highly significant difference.

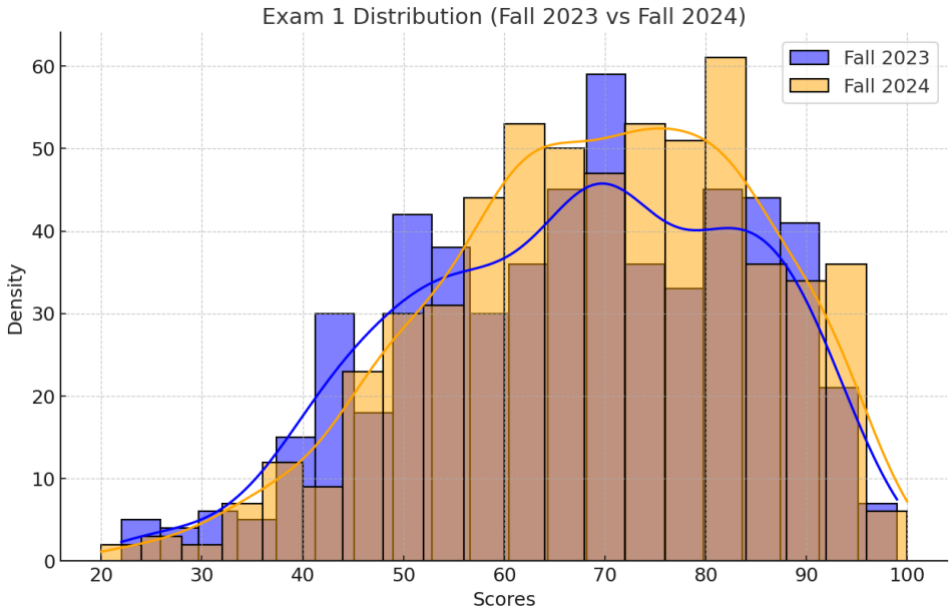

Figure 1: Distribution of Exam 1 comparing Fall 2023 to Fall 2024 cohorts.

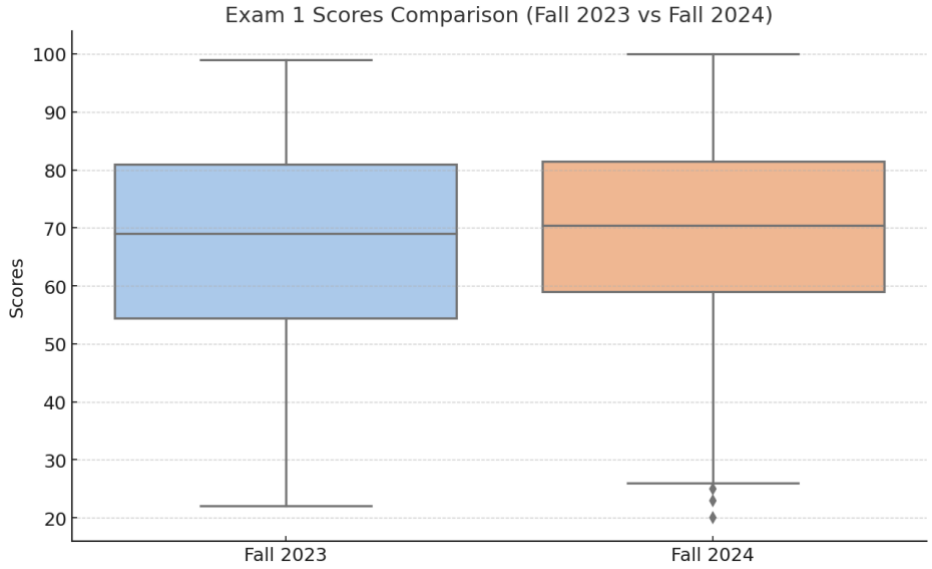

Figure 2: Exam 1 Box and whisker plot comparison between Fall 2023 and Fall 2024

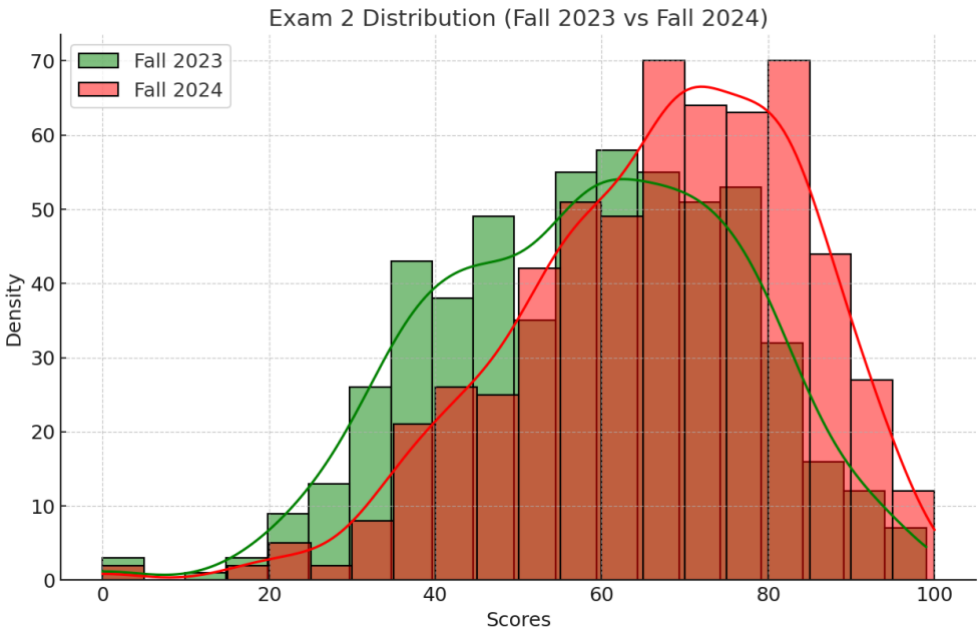

Figure 3: Distribution of Exam 2 comparing Fall 2023 to Fall 2024 cohorts.

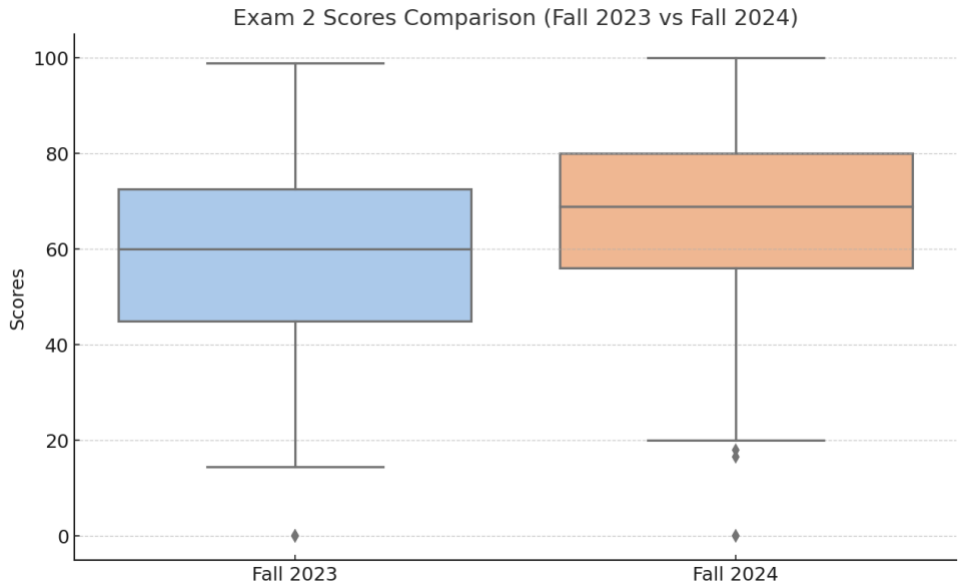

Figure 4: Exam 2 Box and whisker plot comparison between Fall 2023 and Fall 2024

Table 1: Descriptive statistics for each exam

Exam	Exam 1 Fa23	Exam 1 Fa24	Exam 2 Fa23	Exam 2 Fa24
Count	560	590	559	583
Mean	67.20	69.08	58.57	67.02
Median	69	70.5	60	60
Standard Deviation	16.68	15.73	18.06	16.83
2nd Quartile	54.5	59	45	56
3rd Quartile	81	81.5	72.5	80

Academic Dishonesty

As somewhat of a disclaimer, the rates of academic dishonesty are exclusively examined through the lens of lived experiences of the authors this analysis is far from generalizable and is at this point speculative.

The Fall of 2023 resulted in a total of 25 reports of academic misconduct with 7 of those reported students having direct links to academic misconduct on exams. In Fall of 2024 only 1 case was reported for academic misconduct, and it was also related to misconduct on an exam. This is a stark difference in detected dishonest behavior. The types of misconduct seen in the Fall 2023 semester related to exams was largely related to students using outside resources on the exams. These cases included students using content developed prior to the exam as well as the use of generative AI during the exam. While the possibility of a student accessing outside resources during a paper exam is a much further reach than for digital exams, it is considered a non-issue for this format of assessment.

Discussion and Limitations

Exam Comparison

The t-test p-values for both Exam 1 and Exam 2 between Fall 2023 and Fall 2024 suggest that student performance did in fact improve with a paper exam format of assessment. This does counter work by Dendir and Maxwell [10] which suggests no change would exist. There are some limitations with this study that may contribute to this shift such as the limited number of professors included and sharing their student performance for this comparison. Additionally, there were efforts throughout the college to increase the resources available to students through tutoring groups and other out-of-class mechanisms for improving student learning. This could align with the work by Stowell and Bennett [11] which might suggest the efforts made could have reduced the anxiety, but this is hard to substantiate, and is likely a non-issue. In fact, the concept of text anxiety impact is concurrently being studied by colleagues of the authors.

The exams were not identical from semester to semester but were developed by the same course coordinator for both years. The weak evidence of null hypothesis rejection for Exam 1 when compared to the more significant difference for Exam 2 does poise some additional questions

from the authors. At this point there is no conclusive evidence that paper exams resulted in better exam scores over digital exams, although deeper analysis is warranted. What is certain however is that the large volume of students participating in this first-year engineering course eliminates issues with small sample size on a student basis. A comprehensive collection of student performance comparison is under consideration by the authors. Considering that this is a course that is critical for progression through the collection of engineering programs there is currently no process to track the performance of student cohorts longitudinally, although longitudinal studies of student outcome between faculty is a possibility as the college has over 30 faculty participating in teaching this course. Multiple years of course data is available for inquiry, along with the stability of the course coordinator role assignment.

Academic Dishonesty

As mentioned in the results portion of this paper, the academic dishonesty data is far from conclusive, although there is some indication from the disparate difference between rates of reported misconduct between the authors Fall 2023 and Fall 2024 semesters that the transition to paper exams has at least in some ways limited the detectable behavior of cheating. Whether it has conclusively eliminated all academically dishonest behavior is a near impossible feat, but the authors do feel confident that by eliminating the ease of interaction with technological tools such as GenAI for academically dishonest behaviors during examinations is nearly eliminated. In essence bringing back the security of non-fungible student performance. A student behavior that was noted by the authors during the Fall 2024 semester to be markedly different was the mode in which students studied for their exams. Many students spent a significant time working out their practice problems and other study materials by hand on physical media, whether that be paper or a writing enabled tablet. Through discussions with students, it was clear that students were attempting to simulate the exam environment as much as possible. This simulation of assessment environment is far from evidence to suggest the larger difference seen for Exam 2, although it does provide insight for further explorative work.

Conclusions and Future Work

This work was inspired by the striking presence of academic misconduct during exams in a foundational first-year engineering course at TAMU during the Fall 2023 semester. The authors gathered exam data from their own classrooms in Fall 2023 and Fall 2024. Those exams were deidentified before compilation, cleaned to remove outliers of students who dropped the course. This resulted in a total of 560 and 590 students completed Exam 1 for the Fall 2023 and Fall 2024 semesters respectively, with 559 and 583 students completing Exam 2 for Fall 2023 and Fall 2024 respectively. The volume of students allowed for reliable statistical analysis for the scope of this work. Two-sample t-tests indicated significant differences for both exams with a positive increase in student performance. Exam 1 had a much higher p-value of 0.05 suggesting weak evidence for difference between student performance from semester to semester, while the p-value for Exam 2 was very small at 6.9E-16, suggesting a much better student performance for

that exam. The exam difficulty was controlled for as the same course coordinator was responsible for both exams for both years, but there is a possibility of variance between exams that is currently unaccounted for. The argument of improved teacher performance is possible but ignored in this study as at least one of the authors has taught this class over a series of years. Both faculty observed a marked decline in academic dishonest behavior showcased by the drop in reports to the academic dishonesty unit at TAMU from Fall 2023 to Fall 2024 from 25 to 1 reports in total, and 7 reports of academic misconduct on an exam to 1 report from 2023 to 2024 respectively.

This work does warrant deeper study into these phenomena, pulling historical student performance from the entire unit of faculty teaching this course. Additional inquiry into academic dishonesty cases reported is also warranted, although there is some concern that the requests will be denied by the academic dishonesty office, forcing the authors to engage only in self-reporting by colleagues. The authors are also considering a qualitative study to examine perceived exam difficulty through interviewing students who are acting as teaching assistants in later semesters, comparing their experiences as students and assistants along with their own experiences of encountering academically dishonest behavior of their own peers and students.

In conclusion, there does seem to be some difference in student performance suggesting paper exams as a superior mode of academic assessment, although many factors present interesting issues to explore further. Directly assessing academic misconduct rates has been plagued with administrative and bureaucratic issues that have yet to be resolved but is hopefully a point of potential improvement for deeper insight into these issues.

References

[1] C. K. Y. Chan, "Students' perceptions of 'AI-giarism': investigating changes in understandings of academic misconduct," *Education and Information Technologies,* pp. 1-22, 2024.

[2] Z. Li, "Generative AI in Higher Education Academic Assignments: Policy Implications from a Systematic Review of Student and Teacher Perceptions," Massachusetts Institute of Technology, 2024.

[3] K. Shryock, K. Watson, L. White, and T. Balart, "Developing a Model to Assist Faculty with Taming the Next Disruptive Boogeyman [InPress}," *Available at SSRN 4699941,* 2024.

[4] S. Amani *et al.*, "Generative AI Perceptions: A Survey to Measure the Perceptions of Faculty, Staff, and Students on Generative AI Tools in Academia," *arXiv preprint arXiv:2304.14415,* 2023.

[5] L. White, T. Balart, S. Amani, K. J. Shryock, and K. L. Watson, "A preliminary exploration of the disruption of a generative ai systems: Faculty/staff and student perceptions of chatgpt and its capability of completing undergraduate engineering coursework," *arXiv preprint arXiv:2403.01538,* 2024.

[6] D. L. McCabe and L. K. Trevino, "Academic dishonesty: Honor codes and other contextual influences," *The journal of higher education,* vol. 64, no. 5, pp. 522-538, 1993.
[7] P. M. Newton and K. Essex, "How common is cheating in online exams and did it increase during the COVID-19 pandemic? A systematic review," *Journal of Academic Ethics,* vol. 22, no. 2, pp. 323-343, 2024.
[8] T. B. Gallant, "Academic Integrity in the Twenty-First Century: A Teaching and Learning Imperative. ASHE Higher Education Report, Volume 33, Number 5," *ASHE higher education report,* vol. 33, no. 5, pp. 1-143, 2008.
[9] O. L. Holden, M. E. Norris, and V. A. Kuhlmeier, "Academic integrity in online assessment: A research review," in *Frontiers in Education*, 2021, vol. 6: Frontiers Media SA, p. 639814.
[10] S. Dendir and R. S. Maxwell, "Cheating in online courses: Evidence from online proctoring," *Computers in Human Behavior Reports,* vol. 2, p. 100033, 2020.
[11] J. R. Stowell and D. Bennett, "Effects of online testing on student exam performance and test anxiety," *Journal of Educational Computing Research,* vol. 42, no. 2, pp. 161-171, 2010.