Impact of evaluation method shifts on student performance: an analysis of irregular improvement in passing percentages during COVID-19 at an Ecuadorian institution

Esteban Guevara Hidalgo[1*]

*Correspondence:
esteban.guevara@epn.edu.ec

[1] Departamento de Formación
Básica, Escuela Politécnica
Nacional, Ladrón de Guevara
E11-253, Quito, Ecuador

Abstract

The COVID-19 pandemic had a profound impact on education, forcing many teachers and students who were not used to online education to adapt to an unanticipated reality by improvising new teaching and learning methods. Within the realm of virtual education, the evaluation methods underwent a transformation, with some assessments shifting towards multiple-choice tests while others attempted to replicate traditional pen-and-paper exams. This study conducts a comparative analysis of these two types of evaluations, utilizing real data from a virtual semester during the COVID-19 pandemic at an Ecuadorian institution. It aims to assess the impact of transitioning from one evaluation method to the other, revealing fundamental structural differences. These differences can lead to disparities that unfairly advantage or disadvantage certain student groups based on the evaluation method used. Beyond identifying the causes of these discrepancies, the study reveals that, for the specific case and dataset analyzed, the shift to virtual education led to a significant and abrupt increase in passing percentages. Moreover, under one specific type of evaluation, there is a possibility that a minimum of 21.1% of students may have passed a course due to cheating or other forms of academic dishonesty, while at least 5.5% could have failed that course despite possessing the necessary capabilities.

Keywords: Academic integrity, Cheating, Virtual education, Online learning, Online assessments, Education in Ecuador

Introduction

Academic integrity (Barnes 1904; Bretag 2016; 'Teddi' Fishman 2016; Macfarlane et al. 2014; McCabe and Trevino 1993; East and Donnelly 2012; McCabe and Pavela 2004; Lancaster 2021) serves as the bedrock of education. Educational institutions have enshrined this fundamental principle within their charters and codes of honor (McCabe and Trevino 1993; McCabe and Pavela 2004; Whitley Jr and Keith-Spiegel 2001). In some cases, severe penalties are imposed on those found in violation, as the essence of the educational process hinges upon it. The issue of cheating has long been a

concern (Barnes 1904; McCabe et al. 2001; Amigud and Lancaster 2019; Colnerud and Rosander 2009), prompting the development and refinement of various control mechanisms over time (Cizek 1999). These mechanisms can vary significantly depending on the institution and region, often encompassing strict regulations governing items allowed during an exam, dress codes, seating arrangements, and even subtle gestures among students. However, instances of academic dishonesty have increased significantly during the COVID-19 pandemic, due to the shift to virtual education (Ndovela and Marimuthu 2022; Lopez and Solano 2021; Noorbehbahani et al. 2022; Bilen and Matros 2021; Janke et al. 2021; Holden et al. 2021; Dendir and Maxwell 2020; Hill et al. 2021; Newton and Essex 2024). Many efforts have been made to adapt traditional control mechanisms to the virtual environment (Holden et al. 2021; Bilen and Matros 2021; Hylton et al. 2016; Northcutt et al. 2016; Clark et al. 2020), often utilizing platforms and technologies tailored for this purpose (Holden et al. 2021; Hussein et al. 2022, 2020; Ruipérez-Valiente et al. 2021; Du et al. 2022). However, due to cost considerations, many teachers opted for the most accessible and minimal forms of monitoring online exams, such as just requiring students to turn on their cameras (Hylton et al. 2016). This made it challenging to effectively prevent cheating (Noorbehbahani et al. 2022; Newton and Essex 2024), as these methods do not easily verify the student's identity, ensure that no notes are visible, or confirm that no one else is assisting the student (Labayen et al. 2021).

In most cases, evaluations were reduced to multiple-choice question (MCQ) tests, where answers could be easily shared via social networks or instant messaging apps, thereby facilitating cheating (Lancaster 2019; Amigud and Lancaster 2020; Lancaster and Cotarlan 2021). To counter the ease with which answers could be shared, alternative assessment methods emerged (Asgari et al. 2021). For instance, some educators attempted to replicate the format of traditional pen-and-paper exams in a virtual environment. These assessments involved generating unpublished and unique problems crafted by the instructor, ensuring they were not available elsewhere. The focus extended beyond merely providing answers, emphasizing the evaluation of the problem-solving process. However, despite these concerted efforts to deter dishonest behavior, instances of cheating were still identified.

Education in pandemic time

The core issues in education, such as teaching methods and the assessment of student knowledge, have been persistent challenges that educators have continually addressed. Although various approaches have been implemented over the years, no universal solution exists, and outdated and ineffective methods still persist (León and García-Martínez 2021). Recognizing the limitations of traditional methods, modern education incorporated technology into the teaching process to address challenges such as waning interest, short attention spans, and a society deeply immersed in cyberculture (Watson and Tinsley 2013). In a world where tablets, computers, smartphones, messaging apps, social networks, and YouTube have become integral to daily life, it is nearly impossible to envision education without them (Sage et al. 2021). However, the successful adoption of these technologies was gradual for some and abrupt for many due to the COVID-19 pandemic. Below is a brief overview of how lectures and evaluations were conducted during the pandemic within an Ecuadorian institution, highlighting some of the advantages and

disadvantages that emerged. While this overview is based on observations from Ecuador, similar challenges and adaptations have likely been experienced in other countries and institutions as well.

Teaching in pandemic time

Virtual education during the pandemic typically involved the presentation of extensive slides (León and García-Martínez 2021; Levasseur and Sawyer 2006), with a few exceptions. While this approach was common in fields like social sciences before the pandemic, it underwent a significant extension to engineering and science during this period, encompassing important activities such as labs and exercises (Asgari et al. 2021). In many instances, the virtual classroom lacked meaningful interaction between students and instructors, essentially becoming a monologue where teachers read and advanced through slides (Hortsch and Rompolski 2023). Some efforts were made to incorporate new strategies that have emerged in recent years in both virtual and traditional education contexts (Ahshan 2021).

These strategies included the flipped classroom (Tucker 2012; Akçayır and Akçayır 2018; Gilboy et al. 2015) and gamification (Deterding et al. 2011; Dichev and Dicheva 2017; Hamari et al. 2014; Seaborn and Fels 2015; Sailer and Homner 2020), which introduced elements like videos, presentations, crosswords, quizzes, quests, and short tests into lectures. While the flipped classroom has demonstrated success in non-virtual settings (Tucker 2012; Akçayır and Akçayır 2018; Gilboy et al. 2015), one of its key advantages, providing instructors with more available time for interactive engagement with students, has often been underutilized especially in virtual settings. This underutilization may stem from the challenge of adapting face-to-face pedagogical approaches to online environments. Additionally, although students can watch lecture videos at their own pace, the shift to online learning can lead to a different type of workload, as students must manage various activities and assignments remotely (Al-Kumaim et al. 2021). The perceived increase in workload may be less about the volume of tasks and more about the adjustment to new learning methods and the need for instructors to develop appropriate online pedagogical strategies (Hortsch and Rompolski 2023).

In some instances, lectures adopted a rudimentary structure resembling that of a Massive Open Online Course (MOOC) (Bax et al. 2018; Kaplan and Haenlein 2016; Wang and Zhu 2019). Course materials and activities were uploaded to a platform accessible to students at their convenience. Typically, these materials comprised course slides, videos, and supplementary resources such as extended readings. Evaluation in these courses typically relied on projects, crosswords, games, puzzles, and multiple-choice question (MCQ) tests, many of which were machine-marked. These assessments often involved minimal interaction with the instructor and had limited control mechanisms in place.

To enhance interaction and simulate traditional lectures, electronic pens emerged as a valuable tool during the pandemic (Asgari et al. 2021). This technology enabled instructors to engage in real-time writing and drawing on a virtual surface, which could be presented to students through online platforms or shared screens. This allowed instructors to interact with digital material, create diagrams, and solve problems electronically, offering a more interactive experience.

One of the major benefits that emerged from virtual education was the availability of recorded lectures (Nkomo and Daniel 2021). The possibility of reviewing a topic as many times as needed and at one's own pace is an advantage that, in general, was not widely available before the pandemic. However, the availability of these recorded videos and materials after live lectures has somewhat diminished the necessity of attending them, especially when the content mirrors what the teacher covers during synchronous sessions (Levasseur and Sawyer 2006). Moreover, platforms like YouTube often offer a more engaging and comprehensive learning experience compared to traditional lectures (Shoufan 2019). As a result, educators face the challenge of leveraging these new methods, resources, and tools to capture and sustain student attention and interest.

In summary, while virtual education brought notable benefits such as recorded lectures and flexible learning, it also introduced significant challenges (Hortsch and Rompolski 2023). Beyond the benefits and the drawbacks presented above, all these virtual education approaches share common issues: academic dishonesty, student work overload, and reduced interaction between teachers and students, as well as among students themselves.

Evaluations in pandemic time

The course scores were distributed across various activities, such as short tests, projects, homework, videos, posters, and other brief assignments like games, quizzes, and crosswords. One of the most commonly used methods to assess students during the emergency was the Multiple Choice Question (MCQ) virtual tests (Asgari et al. 2021), referred to as **v-Tests** hereafter. In these evaluations, problems were randomly selected from a question bank, and students either chose from several options or entered their answers in a provided box. This signified a major change in fields like engineering and science as the process itself was not evaluated; only the final answer. Consequently, there was no distinction between a student not knowing the answer and making minor arithmetic errors. As an alternative, some instructors within these fields, attempted to replicate traditional pen-and-paper exams in a virtual environment. These evaluations were based on problem-solving, where both the answer and the problem-solving process were graded **(p-Tests)**.

This paper conducts a comparative analysis on these two types of exams, utilizing real data from a virtual semester during the COVID-19 pandemic to assess the impact of transitioning from p-Tests to v-Tests. The study focuses on the potential disparities in student outcomes based on the type of assessment method used and investigates the conditions under which these disparities arise. Additionally, three distinct scenarios are presented to illustrate how certain groups of students may be unfairly advantaged or disadvantaged by different evaluation methods.

Proctoring in pandemic time

With the lockdown and the impossibility of in-person meetings, the urgency of maintaining academic integrity in online assessments became a major concern. In response, three types of remote proctoring mechanisms were implemented:

1. Live proctoring, where a person monitors the examination by watching the students live during an online meeting (Mitra and Gofman 2016; Patael et al. 2022; Hylton et al. 2016).
2. Recorded proctoring, in which the examinee is video recorded, and the recording is reviewed by a human proctor at a later time to assess the integrity of the exam (Hussein et al. 2020).
3. Automated proctoring, where a proctoring system monitors the examination. This system uses statistical methods (Awad Ahmed et al. 2021; Duhaim et al. 2021), artificial intelligence (Chou 2021; Hussein et al. 2022; Nigam et al. 2021), deep learning algorithms (Tiong and Lee 2021), or other techniques (Atoum et al. 2017; Turani et al. 2020; Masud et al. 2022) to identify signals of possible fraud or cheating. A human proctor then reviews these alerts to determine if any misconduct has occurred.

Since the first mechanism is the easiest and most direct to implement, this was the proctoring method used at the institution analyzed in this paper. However, we will examine two different approaches to live proctoring and their potential impact on the integrity of the examination.

Methodology

During the (virtual) fall semester of 2021, a group of students within an Ecuadorian institution unexpectedly underwent an abrupt and unplanned change in their evaluation format. In the first part of the semester (referred to as B_1), they were assessed through two **p-Tests**, while in the second part (B_2), they faced two **v-Tests** (with a small p-Test in between the two v-Tests). This paper aims to compare the results obtained from these 109 first-year engineering students to measure the differences between these two types of tests. It is important to note that this study did not require review and approval from the institution's ethics committee, nor did it involve participant consent, as it consists in a direct analysis of the data from that semester.

Tests descriptions and examination settings

First half of the semester: procedure-graded tests (p-Tests)

Each p-Test had a duration of 1 hour and consisted of four unpublished problems. To ensure a smooth exam experience, students were instructed to access the virtual meeting 15 minutes prior to the exam to mitigate potential issues such as software updates or computer reboots. Subsequently, the teacher conducted a location check, which had been previously communicated during lectures and via email, outlining the specific requirements for how and where they should be situated during the exam. Students were required to manually adjust their standard cameras in a way that allowed the proctor to view not only their faces but also their screens, hands, and desks, without the use of any specialized equipment. This verification process took an average of 30 minutes, after which the exam content was projected/shared on the students' screens. Once the exam commenced, students were prohibited from using the keyboard, mouse, or smartphones. They were required to solve the four problems "by hand" on sheets of paper, and one hour later, scan the entire exercise, including their step-by-step solutions, using a mobile app. These scanned copies were then sent to

the teacher's email and uploaded to the platform. Although this process typically took about 5 minutes, students were allotted 10 minutes for submitting the exam in PDF format. The grading for the p-Tests involved the utilization of electronic pens and evaluated the entire procedure, not just the final answers, which meant that minor arithmetic errors had a minimal impact on the final grade.

Second half of the semester: multiple choice questions tests (v-Tests)

Each v-Test had a duration of 50 minutes and comprised five multiple-choice problems. The platform generated a unique exam for each student, randomly selecting questions from a database containing approximately 20–30 exercises, each offering five possible answers. This database was collaboratively created by seven course teachers. The questions used in the v-Tests did not necessarily have to be entirely new; they could also be modified versions of exercises from the homework assignments. The platform was configured to ensure that the difficulty level of the v-Tests remained consistent for every student.

These v-Tests were administered to 795 students (the total number of students taking Course X during the fall semester of 2021), including the 109 students who had previously been evaluated using p-Tests during the B_1 phase. To mitigate potential network and platform issues, the students were divided into two groups. The first half of students took the exam at a specified time, followed by the second group one hour later, with both groups receiving questions from the exact same database. While students had been instructed on how to position themselves during the test, the settings for the v-Tests did not permit location monitoring prior to the exams. Consequently, many of the 795 students took these v-Tests with minimal supervision. The exams were conducted as online meetings where students were required to have their computer cameras on, but there was no formal remote invigilation system in place. Thus, while the cameras provided a view of the student's faces, there was no dedicated monitoring of their actions or environment beyond ensuring they were visible on camera. During the v-Tests, students were prohibited from using the keyboard, mouse, smartphones, or notes, although it was not always feasible to verify compliance with these restrictions. The platform automatically graded the exams, considering only the final answer and not the process or how that answer was obtained. As a result, minor arithmetic mistakes could significantly impact the final scores.

Results

The p-grade is defined as the average of the p-tests, and similarly, the v-grade is calculated as the average of the v-tests (both graded on a 10-point scale). These grades measure the students' performance during each type of evaluation. The final grade for each student (out of 20 points) is the sum of the p-grade and the v-grade. These quantities were compared and analyzed, and the results are presented below. Although the course score included other activities such as labs and homework, the analysis presented in this paper focuses solely on test performance.

Students performance redistribution during the v-Tests

In Fig. 1, the p- and v-grades are presented in ascending order for each student. For the p-Tests, the students achieved average results ranging from 0 to 9 points, while for the v-Tests, the average scores fell within the range of 0 to 7.5 points, both scores out of 10. Since the v-Tests are multiple-choice exams, the v-grades can only assume specific values, resulting in a stair-like pattern in the data, as depicted in Fig. 1b. Based on the results obtained in the p-Tests, the students were categorized into three groups (see Fig. 1a): Group 1 (G_1) consists of students with the lowest performance, representing the bottom third, scoring between 0 and 2.22. Group 2 (G_2) encompasses students scoring between 2.22 and 5.475, and Group 3 (G_3) comprises the top third of students who achieved the best results.

While the average performance remained similar regardless of the evaluation type (approximately $\bar{p}_{grade} \approx 3.87$ and $\bar{v}_{grade} \approx 3.96$), a noticeable shift in student performance occurred during the second part of the semester when the evaluation method switched to v-Tests. The redistribution of students is illustrated in Fig. 1b, using the same color scheme as the group classification based on the p-grades. This shift in student performance is visually evident in Fig. 2. Notably, students who received the lowest grades in the p-Tests presented significant improvement in the multiple-choice tests, as depicted in Fig. 2a. Conversely, students who achieved higher p-Tests grades experienced a decline in their performance during the v-Tests, as shown in Fig. 2c. With the exception of two out of 36 G_1 students, the rest showed an increase in their performance ranging from 0.34 to 6.24 points during the v-Tests. In contrast, the majority of G_3 students (except one out of 36) experienced a decline in their performance, with reductions ranging from 0.03 to 6.33 points during the v-Tests.

Trends in passing percentages: pre, during and post COVID-19

Before COVID-19, Course X utilized pen-and-paper exams, where students were required to solve problems by hand. These exams were common for all first-semester students (around 800) and were conducted in large auditoriums, with face-to-face invigilation by several supervisors. However, with the onset of the pandemic, the course's evaluation method shifted entirely to virtual MCQ tests, with the sole exception of the case analyzed in this paper (specified in the Methodology section and occurring during Semester S_8). Figure 3 presents actual data of the passing percentages for Course X before, during and after COVID-19. The average passing percentage before (38.37%) and during (75.91%) the pandemic are presented as continuous lines. The transition to virtual education resulted in a noticeable and abrupt increment in the passing percentages (an average of 37.54%).

With the conclusion of the pandemic, the course faced three alternatives regarding evaluations: 1) reverting to the pre-pandemic evaluation methods, completely discarding v-Tests, 2) predominantly assessing the course through v-Tests, or 3) adopting a combination of both evaluation methods. Course X chose the first option, reverting to p-Tests in Semester S_{10}. Consequently, the passing percentage of the course dropped again, reaching values similar to those before the pandemic. However, when the course was evaluated with v-Tests again in Semester S_{11}, its passing percentage returned to a
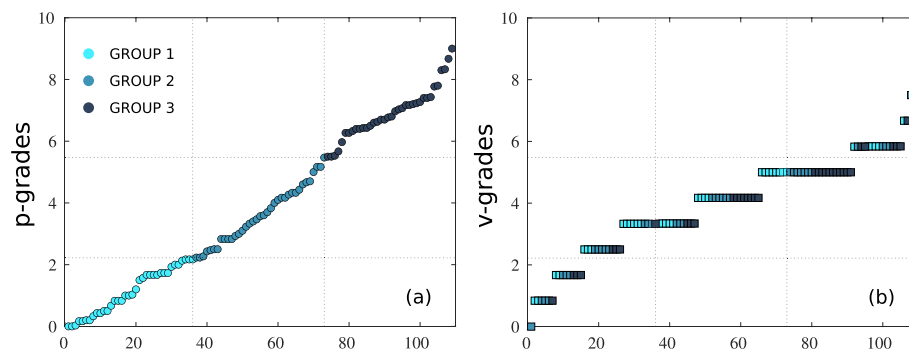
Fig. 1 Grades in ascending order for each of the 109 students. **a** Presents the average of the p-Tests grades (p-grades), while (**b**) shows the average of the v-Tests grades (v-grades). Students were classified into three groups based on their p-grades: Group 1 (G_1: bottom third of students with the lowest results), Group 2 (G_2: middle group), and Group 3 (G_3: top third with the best results). This distribution changed during the second part of the semester when the type of evaluation changed to v-Tests, as depicted in (**b**) using the same color scheme as in (**a**)
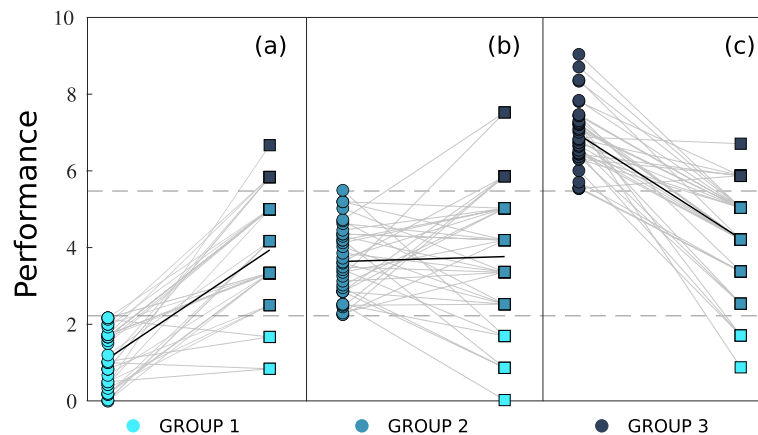

Fig. 2 Change in student performance when transitioning from p-Tests to v-Tests. The figure illustrates the average results obtained by each student during the first part of the semester when assessed with p-Tests (represented by circles) and during the second part when evaluated with v-Tests (represented by squares). Gray lines indicate the extent to which their results improved or diminished. **a** Students belonging to Group 1 (those with the lowest p-Test results) experienced an average improvement of 2.83 points, with some showing up to 6.24 points enhancement during the multiple-choice virtual tests (v-Tests). Conversely, **c** students from Group 3 (those with the highest p-Test results) saw an average decline of 2.71 points, with some experiencing up to 6.33 points reduction

level comparable to those obtained during the pandemic. This suggests that student performance may be influenced by the type of evaluation employed. The next section will attempt to estimate how the final outcome would have differed (during Semester S_8) and how many students might have passed or failed the course if the evaluation had been based solely on p-Tests or v-Tests.

Comparing student outcomes in different evaluation scenarios

Apart from the real-world case described above, which will be referred to as **Scenario 1 (M)**, two other hypothetical scenarios are also considered. **Scenario 2 (P)** assumes the students were evaluated only with p-Tests, and **Scenario 3 (V)** only with v-Tests, both created from the actual data of the v- and p-grades from Semester 8. Figure 5 in
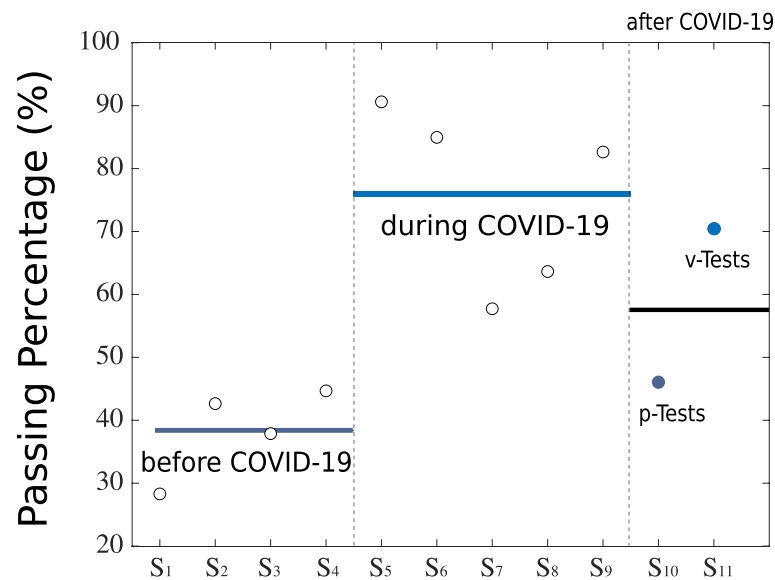
Fig. 3 Passing percentage for Course X across different semesters: before, during, and after the COVID-19 pandemic. Each semester's passing percentage is depicted with circles, and their averages are represented by continuous blue lines. The shift to virtual education marked a significant change in the passing percentage, reflecting an improvement of 35.87%. During the COVID-19 pandemic, apart from the 109 students from Semester S_8, approximately 800 students each semester underwent evaluations using exclusively v-Tests. Post-pandemic, Course X returned to traditional evaluations (p-Tests), leading to a decline in the passing percentage to values akin to those pre-pandemic. Intriguingly, in Semester S_{11}, when the course reverted to v-Tests, the Passing Percentage rebounded to levels comparable to those observed during the pandemic

Appendix compares these scenarios by displaying the final grades Y (over 20) in ascending order for each student. The figure's colors correspond to the original group classification. Vertical dashed lines define three zones to determine a student's course outcome: Zone 1 ($Y < 5$) indicates a failing grade, Zone 2 ($5 \leq Y < 10$) necessitates an additional exam, and Zone 3 ($Y \geq 10$) signifies a passing grade. The zone composition for each scenario is detailed in Tables 1, 2 and 3, where the percentages of students from each group present in each zone are provided, along with the percentage of students failing or passing the course. Although Scenario (V) yields the best results with higher passing (40.37%) and lower failing (13.76%) percentages (Table 3), a more in-depth analysis of the zone composition and how the migration between scenarios and zones occur is discussed below.

Table 1 Scenario 1 (M): p-Tests and v-Tests (Fig. 5a in Appendix)

	Zone 1			Zone 2			Zone 3		
	G_1	G_2	G_3	G_1	G_2	G_3	G_1	G_2	G_3
Zone Composition[a]	15	5	0	21	29	10	0	3	26
	75%	25%	0%	35%	48.33%	16.67%	0%	10.34%	89.66%
		Fail			Exam			Pass	
Total[a]		20			60			29	
		18.35%			55.05%			26.6%	
Group Contribution	13.76%	4.59%	0%	19.27%	26.61%	9.17%	0%	2.75%	23.85%

[a] Short division

Table 2 Scenario 2 (P): Only p-Tests (Fig. 5b in Appendix)

	Zone 1			Zone 2			Zone 3		
	G_1	G_2	G_3	G_1	G_2	G_3	G_1	G_2	G_3
Zone Composition[a]	36	5	0	0	28	0	0	4	36
	87.8%	12.2%	0%	0%	100%	0%	0%	10%	90%
		Fail			Exam			Pass	
Total[a]		41			28			40	
		37.61%			25.69%			36.7%	
Group Contribution	33.03%	4.59%	0%	0%	25.69%	0%	0%	3.68%	33.02%

[a] Short division

Table 3 Scenario 3 (V): Only v-Tests (Fig. 5c in Appendix)

	Zone 1			Zone 2			Zone 3		
	G_1	G_2	G_3	G_1	G_2	G_3	G_1	G_2	G_3
Zone Composition[a]	4	7	4	18	17	15	14	13	17
	26.67%	46.66%	26.67%	36%	34%	30%	31.82%	29.54%	38.64%
		Fail			Exam			Pass	
Total[a]		15			50			44	
		13.76%			45.87%			40.37%	
Group Contribution	3.67%	6.42%	3.67%	16.51%	15.6%	13.76%	12.84%	11.93%	15.6%

[a] Short division

Migration and irregular pass-fail patterns

Figure 4 illustrates the zone composition utilizing data from Tables 2 and 3. The number of students in a specific zone in Scenario 2 (P), denoted as n_P, is shown on the left side of Fig. 4, while the number of students in a particular zone in Scenario 3 (V), denoted a n_V, is displayed on the right. The color gradient reflects student performance, ranging from the lowest (clear blue) to the highest (dark blue). The subzones 1, 2, or 3 are related to the group composition of each zone. For example, in Scenario 3 (P), 36 students from Group 1 and 5 students from Group 2 fail the course (Zone 1), while no students from Group 3 fail the course. The figure allows us to observe how the zones are composed in each scenario, specifically which students are failing or passing the course in each case. $\Delta n = n_V - n_P$ then quantifies the number of students who have transitioned from one zone i and group j in Scenario (P), denoted as Z_{ij}^P, to another zone and group under the virtual scenario (V), represented by Z_{ij}^V. This zone variation Δn is visually represented by circles in Fig. 4, while the migration of students to other zones is depicted with arrows.

Out of the 32 students with the lowest p-Test performances (Z_{11}^P), 18 of them migrated to Zone 2, and 14 to Zone 3 within a virtual scenario. Additionally, 11 students moved from Z_{22}^P to Zone 1 (2 students) and to Zone 3 (9 students), while 19 students with the best p-Tests results, Z_{33}^P, migrated to Zone 1 (4 students) and to Zone 2 (15 students). In a similar comparative analysis between Scenario 1 (M) and Scenario 2 (P), there is no migration to direct passing or failing. Instead, 21 students from Z_{11}^P and 11 from Zone 3 have to take an additional exam. After this exam, only one Z_{32}^P student could be failing the course unfairly. There is no apparent irregular passing.
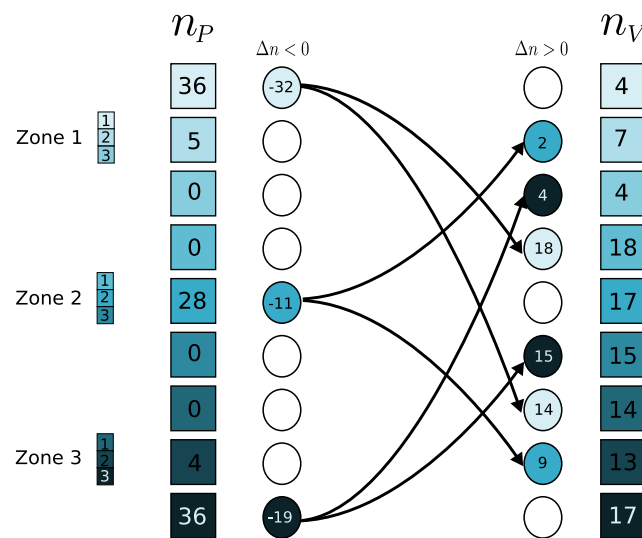
Fig. 4 Zone composition for two different hypothetical scenarios. Scenario 2 (P) assumes the students were evaluated only with p-Tests, and Scenario 3 (V) only with v-Tests. Each zone determines if a student, fails (Z_1), requires an additional exam (Z_2) or passes the course (Z_3). The sub-zones 1, 2, or 3 are related to the group composition of each zone. The number of students, denoted as n_P or n_V, in a specific zone is presented inside boxes. The variation (shown in circles) $\Delta n = n_V - n_P$, measures the number of students who have moved from a zone i and group j in (P), Z_{ij}^P, to another zone and group during the virtual scenario (V), Z_{ij}^V. Before the additional exam, at least a 21.1% (14 students from Z_{11}^P and 9 students from Z_{22}^P) of the students maybe passing irregularly the course. On the other hand, at least 5.5% of students (4 from Z_{33}^P and 2 from Z_{22}^P) could be failing the course despite their actual capabilities. After the exam, these percentages increased to 25.47% (66.67% belongs to G_1) and to 11.92% (84.62% are G_3-students), respectively

Discussion

When we compare the results obtained by students in p-tests with those from v-tests, a noticeable shift in student performance can be observed. Strikingly, students who received the lowest grades in the p-Tests demonstrated a significant improvement in the multiple-choice virtual tests (Fig. 2a). Conversely, students who achieved higher grades in p-Tests experienced a decline in their performance during the v-Tests (Fig. 2c). A decrease in performance when students are evaluated with MCQ tests can be expected, as the procedure is not graded. However, the improvement among students with the lowest performance is particularly noteworthy.

While the availability of recorded lectures after virtual sessions provided all students with the opportunity to review the material at their own pace, this factor remained during the semester analyzed consistent across both p-Tests and v-Tests. Therefore, although recorded lectures could have influenced performance during the pandemic (Nkomo and Daniel 2021), it is unlikely that they were the primary factor contributing to the observed differences between the two types of evaluations.

The disparities in results may be attributed to differing control mechanisms employed in each type of evaluation (Hylton et al. 2016; Dendir and Maxwell 2020). As detailed in the Methodology section, v-Tests took place with minimal oversight, in contrast to p-Tests. Additionally, there is a connection between students' performance during lectures and their p-Test results. The highest p-grades were achieved by students who actively participated in lectures, engaged with the material, turned on their cameras, provided correct answers, and demonstrated genuine interest. Conversely, the lowest p-grades were awarded to students

with poor or nonexistent participation, those facing sanctions for cheating (see Fig. 6 in Appendix), individuals who submitted blank exams, or those who did not attend lectures at all. However, the significant improvement observed during the v-Tests for the students with the lowest performance (G_1) raises questions. This observation could suggest that, at least in the specific case considered in this study, the transition to multiple-choice virtual tests (v-Tests) might have unintentionally favored a particular group of students while disadvantaging others, thereby also influencing the course pass rates.

For Course X, the transition to virtual education led to a significant and abrupt increase in passing percentages (Newton and Essex 2024), rising from 38.37% to 75.91%, as illustrated in Fig. 3. Although a similar behavior was also observed in all first year courses, Course X showed the most substantial increase. This shift in percentage of students passing could be attributed not only to the measures for exam supervision (proctoring) (Clark et al. 2020; Dendir and Maxwell 2020; Duhaim et al. 2021; Janke et al. 2021; Masud et al. 2022) and the evaluation methods employed (Asgari et al. 2021) but also to the teaching resources (Orlov et al. 2021; Gopal et al. 2021). In addition to the shift to common virtual tests and the almost lack of proctoring, lectures for Course X were delivered using slides (León and García-Martínez 2021; Levasseur and Sawyer 2006). In contrast, the course with the smallest improvement in passing percentages (6%) employed optical pencils and non-standardized paper-based tests (p-Tests).

Once the virtual lectures ended, and the course reverted to p-Tests in Semester S_{10}, the passing percentage of the course dropped again to values similar to those before the pandemic. Later, it returned to a level comparable to those obtained during the pandemic again in Semester S_{11} when the course was evaluated with v-Tests once more. These findings highlight that student performance may be significantly influenced by the type of evaluation employed. However, it is essential to note that the groups benefiting from each type of evaluation could differ significantly.

This observation was reinforced when we delved deeper into which students were failing or passing the course (Figs. 4 and 5 in Appendix). Although the average final grade does not show major differences between scenarios ($\overline{Y}_M \approx 7.83$, in the real case where students were evaluated with both p-Tests and v-Tests, $\overline{Y}_P \approx 7.74$ with only p-Tests, and $\overline{Y}_V \approx 7.9$ with only v-Tests), the composition of the zones varied significantly. In Scenario 1 (M) and Scenario 2 (P), students failing (Z_1) or passing (Z_3) the course correspond to those with the worst (G_1) or best performance (G_3), as expected (as detailed in Tables 1 and 2). However, for the scenario where students are solely evaluated with v-Tests (V), the zone composition is counter intuitive, consisting of students from every group (as shown in Table 3).

When students are solely assessed with p-Tests, the number of students who fail the course and belong to the bottom third is $n_P = 36$ (see Fig. 4). However, when evaluated exclusively with v-Tests, this number is reduced to $n_V = 4$. The difference $\Delta n = -32$, represents the number of students who would have initially failed the course when evaluated with p-Tests, but not necessarily when evaluated with v-Tests. For instance, 14 of these 32 students would pass the course, when evaluated only with v-Tests. Similarly, another 9 of 11 students belonging to zone Z_{22}^P would pass. This observation raises the possibility that up to 21.1% of students may have passed the course due to irregularities in the evaluation process, such as cheating or other forms of academic dishonesty, as their performance did not correspond to their results. This percentage increases to 25.47% after the additional exam. Notably, 66.67%

of the students who may be passing irregularly belong to G_1, the group with the lowest p-Test results. On the other hand, at least 5.5% of students (4 from Z_{33}^P and 2 from Z_{22}^P) could be failing the course despite their actual capabilities. After the exam, this percentage increases to 11.92%, with 84.62% of them belonging to G_3, the top third of the class. While these percentages could indicate a potential serious issue, they should be interpreted with caution, as they represent a possibility, even if strong (Chirumamilla and Nguyen-Duc 2020; Ndovela and Marimuthu 2022; Lopez and Solano 2021; Noorbehbahani et al. 2022; Bilen and Matros 2021; Janke et al. 2021; Holden et al. 2021; Dendir and Maxwell 2020; Hill et al. 2021; Newton and Essex 2024; Lancaster 2021), rather than a definitive conclusion.

Conclusion

An incorrect interpretation of the passing rates presented in Fig. 3 could lead to the conclusion that education improved during the pandemic. For example, it might mistakenly suggest that v-Tests enable students to achieve better results. However, in this paper, we have examined the data and questioned the reasons behind this improvement. The inadequate control mechanisms employed facilitated cheating (Noorbehbahani et al. 2022; Newton and Essex 2024), as the tests were not conducted in a controlled environment and answers could be easily shared (Chirumamilla and Nguyen-Duc 2020; Lancaster 2019; Amigud and Lancaster 2020; Lancaster and Cotarlan 2021), not to mention the difficulty in verifying the identities of the individuals taking the exam (Labayen et al. 2021). While the data suggests that some students may have benefited from these inadequacies, it's important to treat these results carefully, as they highlight potential scenarios rather than conclusive outcomes.

Moreover, other factors, not fully captured in this study, could also contribute to the observed differences. These may include variations in student motivation (Chiu et al. 2021) and stress levels (Al-Kumaim et al. 2021), digital capabilities (Limniou et al. 2021), differences in access to technological resources (Abu Talib et al. 2021; Korkmaz et al. 2022), the quality of virtual teaching (Gopal et al. 2021), and the impact of the pandemic on students' physical and mental health (Talevi et al. 2020; Wilson et al. 2021; Zhang et al. 2020).

In fields like engineering and science, assessments often utilize p-Tests because the objective is not merely to test memory, but to evaluate skills such as problem-solving and reasoning. Sometimes the tests are even open book, or students are provided with the formulas needed (even in in-person courses), because knowing them is not the important part, but how they use them. This distinction highlights why the transition to v-Tests in these fields was particularly significant. Asking a student something exactly as it appears in their lecture notes is different from posing a question that requires reasoning and analysis, where they must write an argument justifying their answer. The real question should be, what do we want to test, and what is the best way to test those skills. It is essential to keep seeking the best alternatives that not only effectively assess a wide range of student skills but also ensure academic integrity.

While the effectiveness of v-Tests has not been questioned, only the lack of control mechanisms, it would be interesting for future research to compare pen-and-paper exams and MCQ tests under identical supervision and monitoring conditions to determine the most suitable method for assessing a student's knowledge acquisition. Whether a higher passing percentage is indicative of a quality education, where students are learning more and better, remains an open question.
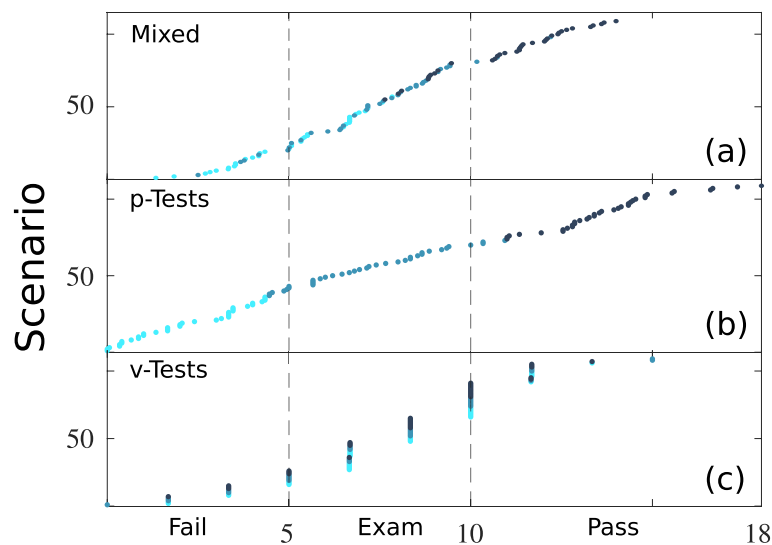
Appendix

Fig. 5 Comparison of Final Grades (over 20) in Ascending Order for Every Student in Three Evaluation Scenarios: **a** Scenario 1 (M): Combining p-Tests and v-Tests (real-world case), **b** Scenario 2 (P): Solely p-Tests, and **c** Scenario 3 (V): Solely v-Tests. Vertical dashed lines demarcate student outcomes into failure (Zone 1), additional exam requirement (Zone 2), and course passage (Zone 3). The color-coding corresponds to the original group classification. Detailed group and zone compositions, along with passing percentages, can be found in Tables 1, 2 and 3
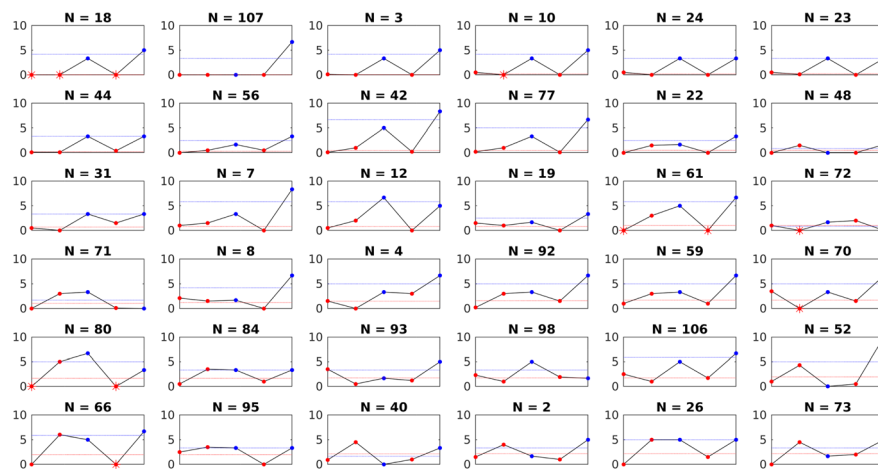

Fig. 6 Individual Performance Analysis of Group G_1. Individual performance of the 36 students with the lowest p-Test performance, which compose group G_1. The p-Tests are shown in red, while the v-Tests are presented in blue. In each subplot, the corresponding p-grade and v-grade are indicated with a dashed line. The figure is organized from the student with the lowest p-grade ($N = 18$) to the highest ($N = 73$). Of the 36 students, 7 were sanctioned for cheating. These students are marked with an asterisk over the corresponding test. For instance, student number 18 was sanctioned for cheating in all p-Tests. Interestingly, no student was sanctioned for cheating in any v-Test. Additionally, two more students sanctioned belonged to G_2, and none of the students sanctioned belonged to G_3. With the exception of 2 out of the 36 students, the rest improved their performance considerably during the v-Tests. Despite this improvement, it is striking to note that, in most cases, performance decreased again during the p-Test between the v-Tests. These results not only supports the observation that students' performance depends on the type of evaluation, but also reinforces it, as performance did not only change from p-Tests to v-Tests, but also from v-Test 1 to p-Test 3, and again to v-Test 2

Abbreviations

MCQs   Multiple choice questions
MOOC   Massive open online course
p-tests   Written procedure virtual tests
v-tests   Multiple choice virtual tests

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to a confidentiality agreement with the institution which participated in the study.

Declarations

Competing interests

The author declare that he has no competing interests.

References

Abu Talib M, Bettayeb AM, Omer RI (2021) Analytical study on the impact of technology in higher education during the age of COVID-19: Systematic literature review. Educ Inf Technol 26(6):6719–6746

Ahshan R (2021) A framework of implementing strategies for active student engagement in remote/online teaching and learning during the covid-19 pandemic. Educ Sci 11(9). https://doi.org/10.3390/educsci11090483

Akçayır G, Akçayır M (2018) The flipped classroom: A review of its advantages and challenges. Comput Educ 126:334–345

Al-Kumaim NH, Alhazmi AK, Mohammed F, Gazem NA, Shabbir MS, Fazea Y (2021) Exploring the impact of the COVID-19 pandemic on university students' learning life: An integrated conceptual motivational model for sustainable and healthy online learning. Sustainability 13(5). https://doi.org/10.3390/su13052546

Amigud A, Lancaster T (2019) 246 reasons to cheat: An analysis of students' reasons for seeking to outsource academic work. Comput Educ 134:98–107. https://doi.org/10.1016/j.compedu.2019.01.017

Amigud A, Lancaster T (2020) I will pay someone to do my assignment: an analysis of market demand for contract cheating services on twitter. Assess Eval High Educ 45(4):541–553

Asgari S, Trajkovic J, Rahmani M, Zhang W, Lo RC, Sciortino A (2021) An observational study of engineering online education during the COVID-19 pandemic. PLoS ONE 16(4):1–17. https://doi.org/10.1371/journal.pone.0250041

Atoum Y, Chen L, Liu AX, Hsu SDH, Liu X (2017) Automated online exam proctoring. IEEE Trans Multimedia 19(7):1609–1624. https://doi.org/10.1109/TMM.2017.2656064

Awad Ahmed FR, Ahmed TE, Saeed RA, Alhumyani H, Abdel-Khalek S, Abu-Zinadah H (2021) Analysis and challenges of robust e-exams performance under COVID-19. Results Phys 23:103987. https://doi.org/10.1016/j.rinp.2021.103987

Barnes E (1904) Student honor: A study in cheating. Int J Ethics 14(4):481–488

Bax S (2018). MOOCs as a new technology: approaches to normalising the MOOC experience for our learners; paper posthumously transcribed by Marina Orsini-Jones. In M. Orsini-Jones & S. Smith (Eds), Flipping the blend through MOOCs, MALL and OIL – new directions in CALL (pp. 9-16). Research-publishing.net. https://doi.org/10.14705/rpnet.2018.23.785

Bilen E, Matros A (2021) Online cheating amid COVID-19. J Econ Behav Organ 182:196–211. https://doi.org/10.1016/j.jebo.2020.12.004

Bretag T (2016) Defining Academic Integrity: International Perspectives – Introduction. Springer Singapore, Singapore, pp 3–5. https://doi.org/10.1007/978-981-287-098-8_76

Chirumamilla S, Nguyen-Duc A (2020) Cheating in e-exams and paper exams: the perceptions of engineering students and teachers in Norway. Assess Eval High Educ 45(7):940–957. https://doi.org/10.1080/02602938.2020.1719975

Chiu TK, Lin TJ, Lonka K (2021) Motivating online learning: The challenges of COVID-19 and beyond. Asia Pac Educ Res 30(3):187–190

Chou T (2021), Apply Explainable AI to Sustain the Assessment of Learning Effectiveness. Complexity, Informatics and Cybernetics: Proceedings of The 12th International Multi-Conference on Complexity, Informatics and Cybernetics (IMCIC 2021), 113. https://www.iiis.org/CDs2021/CD2021Spring/PapersZ2.htm#/

Cizek G (1999) Cheating on Tests: How To Do It, Detect It, and Prevent It. Taylor & Francis. https://books.google.com.ec/books?id=j8qQAgAAQBAJ

Clark TM, Callam CS, Paul NM, Stoltzfus MW, Turner D (2020) Testing in the time of COVID-19: A sudden transition to unproctored online exams. J Chem Educ 97(9):3413–3417. https://doi.org/10.1021/acs.jchemed.0c00546

Colnerud G, Rosander M (2009) Academic dishonesty, ethical norms and learning. Assess Eval High Educ 34(5):505–517

Dendir S, Maxwell RS (2020) Cheating in online courses: Evidence from online proctoring. Comput Hum Behav Rep 2:100033

Deterding S, Dixon D, Khaled R, Nacke L (2011). From Game Design Elements to Gamefulness: Defining Gamification. In: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (pp. 9-15). New York: ACM. https://doi.org/10.1145/2181037.2181040

Dichev C, Dicheva D (2017) Gamifying education: what is known, what is believed and what remains uncertain: a critical review. Int J Educ Technol High Educ 14(1):1–36

Du J, Song Y, An M, An M, Bogart C, Sakr M (2022) Cheating detection in online assessments via timeline analysis. In: Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1. pp 98–104. New York: ACM.

Duhaim AM, Al-Mamory SO, Mahdi MS (2021) Cheating detection in online exams during COVID-19 pandemic using data mining techniques. Webology 19(1):341–366

East J, Donnelly L (2012) Taking responsibility for academic integrity: A collaborative teaching and learning design. J Univ Teach Learn Pract 9(3):2

Gilboy MB, Heinerichs S, Pazzaglia G (2015) Enhancing student engagement using the flipped classroom. J Nutr Educ Behav 47(1):109–114

Gopal R, Singh V, Aggarwal A (2021) Impact of online classes on the satisfaction and performance of students during the pandemic period of COVID 19. Educ Inf Technol 26:6923–6947

Hamari J, Koivisto J, Sarsa H (2014) Does gamification work?–a literature review of empirical studies on gamification. In: 2014 47th Hawaii international conference on system sciences. pp 3025–3034. Piscataway: IEEE.

Hill G, Mason J, Dunn A (2021) Contract cheating: an increasing challenge for global academic community arising from COVID-19. Res Pract Technol Enhanc Learn 16(1):1–20

Holden OL, Norris ME, Kuhlmeier VA (2021) Academic integrity in online assessment: A research review. In Frontiers in Education. Lausanne: Frontiers. p. 258.

Hortsch M, Rompolski K (2023) The freedom to teach (at the best). Anat Sci Educ 16(2):189–195. https://doi.org/10.1002/ase.2240

Hussein F, Al-Ahmad A, El-Salhi S, Alshdaifat E, Al-Hami M (2022) Advances in contextual action recognition: Automatic cheating detection using machine learning techniques. Data 7(9). https://doi.org/10.3390/data7090122

Hussein MJ, Yusuf J, Deb AS, Fong L, Naidu S (2020) An evaluation of online proctoring tools. Open Praxis. https://doi.org/10.5944/openpraxis.12.4.1113

Hylton K, Levy Y, Dringus LP (2016) Utilizing webcam-based proctoring to deter misconduct in online exams. Comput Educ 92:53–63

Janke S, Rudert SC, Petersen Änne, Fritz TM, Daumiller M (2021) Cheating in the wake of COVID-19: How dangerous is ad-hoc online testing for academic integrity? Comput Educ Open 2:100055. https://doi.org/10.1016/j.caeo.2021.100055

Kaplan AM, Haenlein M (2016) Higher education and the digital revolution: About moocs, spocs, social media, and the cookie monster. Bus Horiz 59(4):441–450

Korkmaz Ö, Erer E, Erer D (2022) Internet access and its role on educational inequality during the COVID-19 pandemic. Telecommun Policy 46(5):102353

Labayen M, Vea R, Flórez J, Aginako N, Sierra B (2021) Online student authentication and proctoring system based on multimodal biometrics technology. IEEE Access 9:72398–72411. https://doi.org/10.1109/ACCESS.2021.3079375

Lancaster T (2019) Social media enabled contract cheating. Can Perspect Acad Integr 2(2):7–24

Lancaster T (2021) Academic dishonesty or academic integrity? using natural language processing (nlp) techniques to investigate positive integrity in academic integrity research. J Acad Ethics 19(3):363–383

Lancaster T, Cotarlan C (2021) Contract cheating by stem students through a file sharing website: a COVID-19 pandemic perspective. Int J Educ Integr 17(1):1–16

León SP, García-Martínez I (2021) Impact of the provision of powerpoint slides on learning. Comput Educ 173:104283

Levasseur DG, Sawyer JK (2006) Pedagogy meets powerpoint: A research review of the effects of computer-generated slides in the classroom. Rev Commun 6(1–2):101–123. https://doi.org/10.1080/15358590600763383

Limniou M, Varga-Atkins T, Hands C, Elshamaa M (2021) Learning, student digital capabilities and academic performance over the COVID-19 pandemic. Educ Sci 11(7). https://doi.org/10.3390/educsci11070361

Lopez KM, Solano DM (2021) Ethics of Cheating: Effects of the COVID-19 Pandemic on Academic Honesty, chap 4. pp 63–77. https://doi.org/10.1021/bk-2021-1401.ch004

Macfarlane B, Zhang J, Pun A (2014) Academic integrity: a review of the literature. Stud High Educ 39(2):339–358. https://doi.org/10.1080/03075079.2012.709495

Masud MM, Hayawi K, Mathew SS, Michael T, El Barachi M (2022) Smart online exam proctoring assist for cheating detection. In: Li B, Yue L, Jiang J, Chen W, Li X, Long G, Fang F, Yu H (eds) Advanced Data Mining and Applications. Springer International Publishing, Cham, pp 118–132

McCabe DL, Pavela G (2004) Ten (updated) principles of academic integrity: How faculty can foster student honesty. Chang Mag High Learn 36(3):10–15. https://doi.org/10.1080/00091380409605574

McCabe DL, Trevino LK (1993) Academic dishonesty: Honor codes and other contextual influences. J High Educ 64(5):522–538

McCabe DL, Treviño LK, Butterfield KD (2001) Cheating in academic institutions: A decade of research. Ethics Behav 11(3):219–232

Mitra S, Gofman MI (2016) Towards greater integrity in online exams submission type: Emergent research forum papers. https://api.semanticscholar.org/CorpusID:37618784

Ndovela S, Marimuthu M (2022) Prevalence of online cheating during the covid-19 pandemic. In: Singh UG, Nair CS, Blewett C, Shea T (eds) Academic Voices. Chandos Publishing, pp 443–455. https://doi.org/10.1016/B978-0-323-91185-6.00039-2

Newton PM, Essex K (2024) How common is cheating in online exams and did it increase during the COVID-19 pandemic? A systematic review. J Acad Ethics 22(2):323–343

Nigam A, Pasricha R, Singh T, Churi P (2021) A systematic review on ai-based proctoring systems: Past, present and future. Educ Inf Technol 26(5):6421–6445

Nkomo LM, Daniel BK (2021) Sentiment analysis of student engagement with lecture recording. TechTrends 65(2):213–224

Noorbehbahani F, Mohammadi A, Aminazadeh M (2022) A systematic review of research on cheating in online exams from 2010 to 2021. Educ Inf Technol 27(6):8413–8460.

Northcutt CG, Ho AD, Chuang IL (2016) Detecting and preventing "multiple-account" cheating in massive open online courses. Comput Educ 100:71–80

Orlov G, McKee D, Berry J, Boyle A, DiCiccio T, Ransom T, Rees-Jones A, Stoye J (2021) Learning during the COVID-19 pandemic: It is not who you teach, but how you teach. Econ Lett 202:109812. https://doi.org/10.1016/j.econlet.2021.109812

Patael S, Shamir J, Soffer T, Livne E, Fogel-Grinvald H, Kishon-Rabin L (2022) Remote proctoring: Lessons learned from the COVID-19 pandemic effect on the large scale on-line assessment at Tel Aviv university. J Comput Assist Learn 38(6):1554–1573. https://doi.org/10.1111/jcal.12746

Ruipérez-Valiente JA, Jaramillo-Morillo D, Joksimović S, Kovanović V, Muñoz-Merino PJ, Gašević D (2021) Data-driven detection and characterization of communities of accounts collaborating in moocs. Futur Gener Comput Syst 125:590–603

Sage K, Jackson S, Fox E, Mauer L (2021) The virtual COVID-19 classroom: surveying outcomes, individual differences, and technology use in college students. Smart Learn Environ 8(1):1–20

Sailer M, Homner L (2020) The gamification of learning: A meta-analysis. Educ Psychol Rev 32(1):77–112

Seaborn K, Fels DI (2015) Gamification in theory and action: A survey. Int J Hum Comput Stud 74:14–31

Shoufan A (2019) What motivates university students to like or dislike an educational online video? A sentimental framework. Comput Educ 134:132–144. https://doi.org/10.1016/j.compedu.2019.02.008

Talevi D, Socci V, Carai M, Carnaghi G, Faleri S, Trebbi E, Di Bernardo A, Capelli F, Pacitti F (2020) Mental health outcomes of the COVID-19 pandemic. Riv Psichiatr 55(3):137–144

'Teddi' Fishman T (2016) Academic Integrity as an Educational Concept, Concern, and Movement in US Institutions of Higher Learning. Springer Singapore, Singapore, pp 7–21. https://doi.org/10.1007/978-981-287-098-8_1

Tiong LCO, Lee HJ (2021). E-cheating prevention measures: Detection of cheating at online examinations using deep learning approach – a case study. Retrieved from https://arxiv.org/abs/2101.09841. Accessed 21 Nov 2024.

Tucker B (2012) The flipped classroom: Online instruction at home frees class time for learning. Educ Next 12(1):82–83

Turani AA, Alkhateeb JH, Alsewari AA (2020) Students online exam proctoring: A case study using 360 degree security cameras. In: 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE). pp 1–5. https://doi.org/10.1109/ETCCE51779.2020.9350872

Wang K, Zhu C (2019) Mooc-based flipped learning in higher education: students' participation, experience and learning performance. Int J Educ Technol High Educ 16(1):1–18

Watson D, Tinsley D (2013) Integrating information technology into education. Berlin: Springer.

Whitley BE Jr, Keith-Spiegel P (2001) Academic integrity as an institutional issue. Ethics Behav 11(3):325–342

Wilson OW, Holland KE, Elliott LD, Duffey M, Bopp M (2021) The impact of the COVID-19 pandemic on us college students' physical activity and mental health. J Phys Act Health 18(3):272–278

Zhang Y, Zhang H, Ma X, Di Q (2020) Mental health problems during the COVID-19 pandemics and the mitigation effects of exercise: a longitudinal study of college students in china. Int J Environ Res Public Health 17(10):3722

Publisher's Note