Discover Education

Case Study

Online assessment in the age of artificial intelligence

Alexander Stanoyevitch[1]

Received: 31 October 2023 / Accepted: 18 July 2024
Published online: 19 August 2024

Abstract
Online education, while not a new phenomenon, underwent a monumental shift during the COVID-19 pandemic, pushing educators and students alike into the uncharted waters of full-time digital learning. With this shift came renewed concerns about the integrity of online assessments. Amidst a landscape rapidly being reshaped by online exam/homework assistance platforms, which witnessed soaring stocks as students availed its questionable exam assistance, and the emergence of sophisticated artificial intelligence tools like ChatGPT, the traditional methods of assessment faced unprecedented challenges. This paper presents the results of an observational study, using data from an introductory statistics course taught every semester by the author, and delves into the proliferation of cheating methods. Analyzing exam score results from the pre and post introduction of ChatGPT periods, the research unpacks the extent of cheating and provides strategies to counteract this trend. The findings starkly illustrate significant increases in exam scores from when exams of similar difficulty were administered in person (pre-Covid) versus online. The format, difficulty, and grading of the exams was the same throughout. Although randomized controlled experiments are generally more effective than observational studies, we will indicate when we present the data why experiments would not be feasible for this research. In addition to presenting experimental findings, the paper offers some insights, based on the author's extensive experience, to guide educators in crafting more secure online assessments in this new era, both for courses at the introductory level and more advances courses The results and findings are relevant to introductory courses that can use multiple choice exams in any subject but the recommendations for upper-level courses will be relevant primarily to STEM subjects. The research underscores the pressing need for reinventing assessment techniques to uphold the sanctity of online education.

Keywords Online assessments · ChatGPT · Integrity of exams · Pandemic-induced education · Online cheating · Artificial intelligence in education · Online exam strategies · Assessment challenges · Student behavior

1 Introduction

Distance learning dates back to 1728, when Caleb Phillips advertised in the Boston Globe newspaper his shorthand course offering by mail correspondence for students throughout the United States; see [1]. The internet provided distance learning with a tremendous boost by greatly reducing the turn-around times between professor-student correspondences. The development of new technologies has significantly increased the quality of online courses over the years since. The single most significant catalyst for online education came with the pandemic, where every instructor, from primary education

✉ Alexander Stanoyevitch, alex.stanoyevitch@gmail.com | [1]Department of Mathematics, California State University-Dominguez Hills, 1000 East Victoria Street, Carson, CA 90747, USA.

through graduate school levels, needed to quickly develop sufficient expertise in order to offer online versions of all of their courses that were forced upon them.

The pandemic coincided with, perhaps, the most profound revolution in distance learning ever. Students and instructors who never had any experience with online learning suddenly became full-time users. Debates flourished and studies came out in droves of the effectiveness of online courses. A nice study of the impact of Covid-19 on 309 courses given at eight colleges and universities spanning four continents is given in Bartolic et al. [2].

A crucial concern regarding online courses is the integrity of assessments. Before the pandemic, many instructors resisted the idea of offering their courses online, fearing that assessments would be compromised. The reality is that, compared to administering exams in a classroom setting under the watchful eye of a proctor, allowing students to take exams online at home makes it almost impossible to prevent cheating. There are evident actions students might take during an online exam, with or without permission: opening their books and notes, consulting the web, collaborating with classmates, or seeking assistance from advanced students. Several companies, who ostensibly advertise that they offer help with homework, had no qualms about helping students cheat on exams. During the pandemic, several of these companies and their shareholders profited immensely from this dubious practice; One of them saw their market capitalization nearly quadrupling from the onset of the pandemic to its zenith. For a modest monthly fee, students can obtain one-on-one assistance from a tutor, typically lowly paid from a third-world country, to complete their exams and gain access to an extensive database of solved exam problems across various subjects. A colleague of the author purchased such a subscription from one of these companies to understand the full extent it could undermine online assessment, and the findings were alarming. This instructor discovered that their exam questions, complete with answers, appeared on the platform the very next day!

The focus of this paper concerns assessing student performance in online courses when the exams are administered without any sort of online proctoring. In a face-to-face class, when we proctor an exam in the classroom, we have complete control over what resources students can use. For example, for a math exam, we typically will allow calculators, but neither laptops nor cellphones. If the same exam is given online without supervision, there are many things that cannot be controlled. Despite instructions, students might use their cellphones or computers to access online resources or even network with classmates on the exam questions. Our paper will provide compelling evidence of the prevalence of cheating with online exams using data from the author's introductory large lecture statistics courses, which he has been teaching every fall and spring semesters for over a decade, with 120–150 students each semester.

Another stumbling block in ensuring the integrity of online assessments is the recent emergence of artificial intelligence-driven large language models, most notably ChatGPT, which is owned and hosted by Microsoft. Google has developed its own large language model, known as Gemini. Unlike services that employ tutors and maintain a database of exam questions and solutions, AI-based large language models have been trained on vast data sets and can answer questions from a myriad of contexts with remarkable precision. ChatGPT offers free access to an earlier version (Chat GPT 3.5) with a limit on the number of questions within a certain timeframe. Subscriptions are available that eliminate these restrictions and use a more powerful model (Chat GPT 4).

Figure 1, sourced from OpenAI's paper introducing their ChatGPT version 4, illustrates its performance across a range of well-known exams:

This graphic shows that most all standardized exams cannot be administered without proper proctoring.

In this paper, we explore the present landscape and the future of assessment in online courses, considering third party online homework/exam assistance platforms and ChatGPT. Often, it is neither practical nor possible to proctor exams in an online format. Thus, this issue holds significant importance for the future of online education. Our purpose here will be to provide guidance and advice for giving online proctored exams in the era of large language models such as Chat GPT. Most of this advice stems from the author's extensive experience and findings in using online assessments over the past four years, in mathematics and statistics courses, both at the introductory and advanced undergraduate levels. The key takeaway is that while it remains feasible to conduct online assessments, ensuring the integrity of exams requires a substantial amount of additional effort. This state of affairs will certainly impact university and individual instructor decisions on whether to use online assessment.

2 Review of related literature

For many years before the pandemic, internet-based distance learning has been offered and developed at many universities throughout the world. It has made education more equitable by allowing students to take college classes (and earn degrees) without having to make a major and economically difficult move, waste time with long commutes,

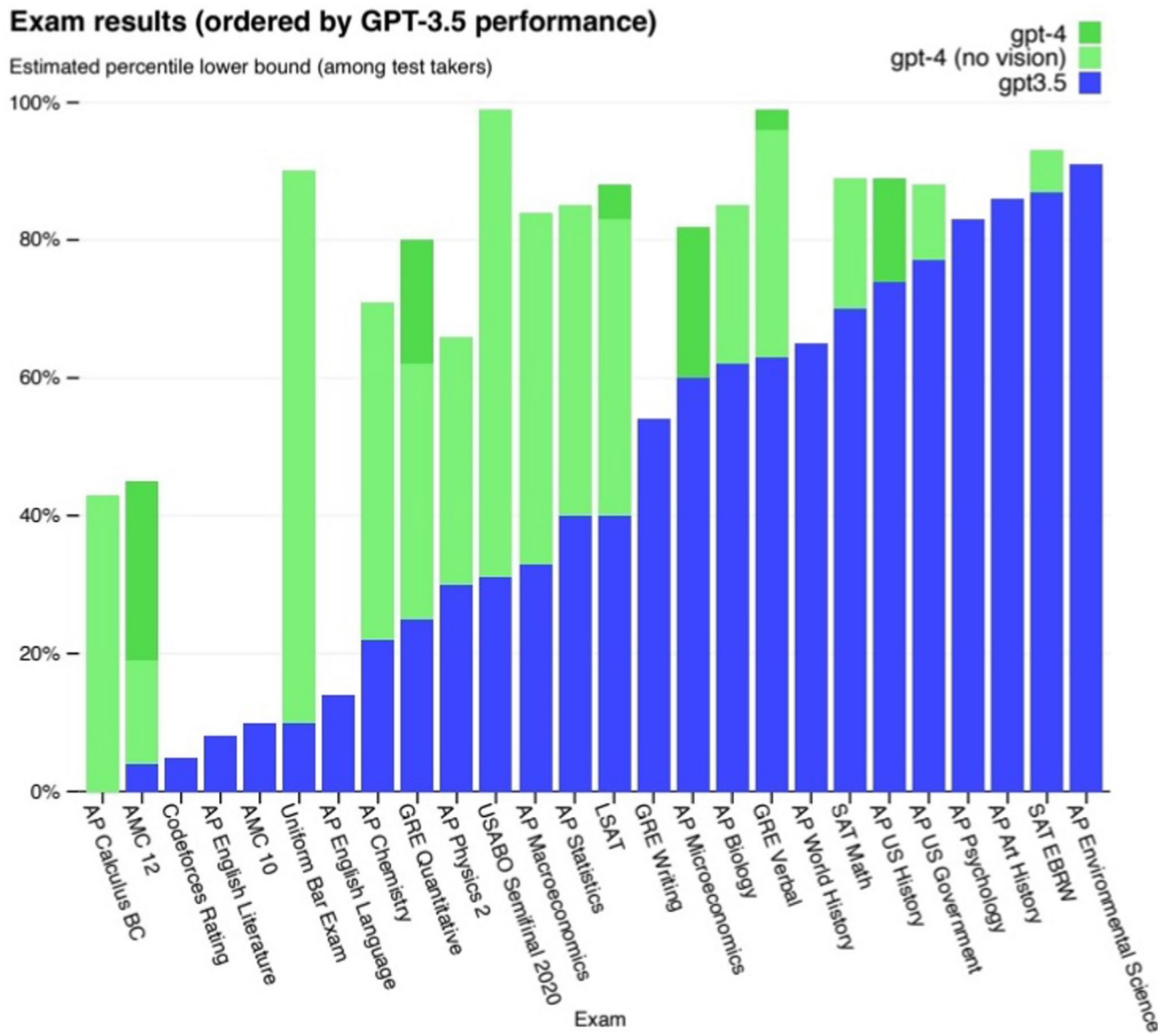

Fig. 1 This figure is taken from the GPT-4 Technical Report by OpenAI (the company behind ChatGPT) accompanying the introduction of ChatGPT-4 in February 2023. The performance percentiles on various tests are depicted: the blue bars represent the previous version, GPT-3.5, and the green bars represent GPT-4. For instance, on the GRE-Verbal exam (a national test used for graduate school admissions), GPT-3.5 scored better than approximately 60% of the students taking the exam, whereas GPT-4 outperformed about 98%

and also to be able to schedule classes around their jobs or family obligations. According to the Oxford Learning College [1], online courses date back to 1965 (before the internet) when the University of Alberta used networked IBM 1500 computers. Also, Massive Open Online Courses (MOOCs) were first offered through MIT in 2012. In 2019, the last year before the pandemic hit, the number of higher education students taking online courses was 36.9%; see [3]. With the increasing prevalence of online education, much research has been done over the years. The meta-study by Bernard et al. [4] examined 232 separate studies of attitudes, achievement and retention outcomes and contains numerous useful research paper references. Horspool and Lange [5] compared online versus face-to-face learning and examined student's perceptions, behaviors, and success rates. Their findings indicate that schedule flexibility is a major factor when students choose an online course over a face-to-face option and that students in both formats felt that they experience high-quality communications with the instructor. They also found that online students study more at home than face-to-face students, but had more limited peer-to-peer communication. The authors present several recommendations, based on their findings, to improve the online course experience. The Horizon 2020 TeSLA Project [6], funded by the European Union, is a valuable resource for online learning. In particular, the Publications

from this project contain a wealth of literature on the assessment topics relating to the topics of this paper. In an effort to assist institutions who are hesitant to adopt online learning course offerings, Fidalgo et al. [7] conducted an extensive survey on undergraduate students in three countries to find out their feelings and concerns about online courses (the survey was done before Covid-19). Shortly into the pandemic, Gammage et al. [8] reviewed assessment security practices in safeguarding academic integrity at several different universities.

Before the onset of COVID-19, the prevalence of online course offerings was shaped by student demand and the willingness of faculty and universities to provide this instructional mode. Online assessment has consistently posed challenges, with difficulties in ensuring assessment integrity discouraging many faculty members from embracing online education. With COVID-19's arrival, faculty found themselves unprepared in the midst of the spring semester (or quarter) of 2020. Traditional face-to-face instruction came to a sudden halt, compelling all instructors to rapidly transition to online teaching. Unfortunately, the outcomes were not always favorable.

Determining the exact extent to which students cheat or even anticipating all possible methods they might employ is challenging. Yet, studies indicate a notable surge in the number of students admitting to cheating on online exams during the COVID-19 pandemic. A recent meta-study by Newton and Essex [9], encompassing 19 studies and 4,672 participants from as far back as 2012, revealed that self-reported online exam cheating rose from a pre-COVID rate of 29.9 to 54.7% during the pandemic. The primary reason students cited for cheating was the mere availability of an opportunity. It's pivotal to recognize that such studies, dependent on voluntary responses, can introduce bias, potentially underestimating the real figures. Nevertheless, the substantial uptick during the pandemic is of significant concern. Noorbehbahani et al. [10] carried out a meta-study reviewing 58 papers on online cheating from 2010 to 2021. Their work aimed to summarize patterns in cheating types, detection methods, and prevention strategies in online environments.

In Dendir and Maxwell [11], compared student scores on the same online exams when some were administered without proctoring and others with online proctoring. They did this comparison in two different courses: Economics and Geography, throughout the semester for a total of 3 high-stakes exams for each class. Their analysis showed significantly higher scores for the non proctored exams over the online proctored exams, over both courses and for all three exams. Many college instructors make use of test banks to create multiple choice exams. Such questions can promptly appear in databases of companies that provide assistance to students on exams and homework. Golden and Kohlbeack [12] did experiments with their online accounting exams and found that if test bank questions are modified by paraphrasing, the student scores on them dropped on average from 80 to 69%. This paper was published before the release of Chat GPT, which can serve as another powerful tool to allow students to cheat on exams.

This paper will showcase the outcomes of experiments designed to gauge student cheating, particularly with the utilization of online student test assistance platforms during the pandemic, in the context of the author's large introductory statistics lecture. The findings indicate that cheating via such platforms was rampant. Each semester, the author typically offers one upper-level course alongside the introductory class, with a smaller enrollment. The specialized nature of the upper-level course makes it harder for students to cheat using such platforms. However, the advent of ChatGPT exposes both course types to novel cheating techniques. This paper will delve into the diverse methods students resort to for cheating and the strategies to mitigate them. Although the primary focus is on mathematics and statistics courses, the insights should be invaluable to instructors across various disciplines.

Holden et al. [13] suggest distinct strategies to curb student cheating in online exams, differing from our propositions. Their expertise lies in non-STEM courses, where they recommend three types of video surveillance techniques and AI-driven plagiarism detection methods. However, video surveillance can be time-consuming, and it may pose ethical and legal challenges. Jia and He [14] at Beijing University crafted an AI-based proctoring system tailored for mathematics courses, utilizing automated video analysis through AI algorithms. They found it to be cost-efficient and highly effective in curbing student cheating in online evaluations.

A strategy closer to our paper's methodology is proposed by Nguyen et al. [15]. In the realm of Chemistry exams, they introduced several economical assessment formats that substantially reduce cheating while still meeting learning objectives.

While large language models can inadvertently assist students in cheating during online exams, they can also augment instruction in numerous beneficial ways, as demonstrated by Jeon and Lee [16].

3 Methods and results

One of the main purposes of this paper is to present evidence showing the extent of students cheating in the author's introductory statistics class when it is offered online compared to when it is offered in the classroom. This is an observational study: the author gives this large lecture statistcs course every fall and spring semester with enrollments in the range 120–150, and has been doing so for the past 10 years. Each semester, there are four exams and a final exam, all multiple choice. Great care it taken to assure that the difficulty of each exam remains the same each semester and also the grading is done completely objectively and simply (each answer gets full credit if correct or no credit if wrong, there is no partial credit or need for any subjective assessment). We collected grade distributions for each exam over all the semesters. Prior to the pandemic, these distributions were quite similar for each exam over the different semesters, but once the pandemic hit and the exams were administered online, the scores increased markedly.

The most reliable statistical methods for collecting data are randomized controlled experiments. But such a design would not be feasible to test our hypothesis. We would need to announce during registration, that students who sign up for this class would be randomly assigned to one of two groups: one group would take their exams in class and the other online (but otherwise the classes would be identical). Few if any students would be willing to sign up for such a class and even if they did, as our research will show, the students taking the exams online would have much higher scores (on the same exams), and this would lead to chaos. Another design might be to offer two different classes: one in which the students would take their exams online and the other where they would take the exams in class. The exams and other aspects of the class would be identical. This design, however, would be flawed, since there would exist many lurking variables and differences stemming from which students choose which format. So a retrospective observational study, as we do in this paper, appears to be the best method for performing such comparisons.

As discussed above, we will present different ways that students have been able to cheat and the prevalence of such cheating using test homework help offered by several companies. We will also explore the new challenges introduced by ChatGPT.

The author has also regularly been teaching an upper-level mathematics course in machine learning every spring semester since before the pandemic. This class typically has an enrollment from 20 to 30. The exams were all take-home exams but very different in nature from the multiple choice introductory statistics exams. The questions involve mixes of mathematical and conceptual reasoning, and some involve computer programming. The recent availability of large language models like Chat GPT, has introduced some problems in assessment. Chat GPT, for example, is capable of writing computer programs. This paper will also discuss ways to prevent students from using such large-language models on exams, and these mitigation methods have been tested during several semesters.

4 Introductory large lecture statistics course

4.1 Part 1A: pre-chat GPT, introductory-level large lecture statistics course

The author's venture into online teaching and assessment began abruptly in the middle of the spring semester of 2020 when the pandemic-induced lockdowns commenced and universities suspended all on-campus classes. All faculty at the author's (California State University) campus were provided with some online teaching tools (including Zoom, Blackboard, and Camtasia) and were given just one week to prepare to continue all of their courses online, for an indefinite amount of time. Initially, our university anticipated this arrangement to last less than a month, an estimate that, in hindsight, proved greatly optimistic.

Prior to the pandemic, the author had essentially no experience with online education. Although there were opportunities to give online courses, the author's reservations with online assessment along with his preference for in-person rather than online interactions deterred any further exploration of this instructional method. The swift shift to online was thrust upon us all, requiring urgent preparation and numerous decisions. The university aimed for maximum flexibility: faculty could conduct their lessons synchronously (live Zoom sessions), asynchronously (pre-recorded lectures), or a mix of both. For live sessions, instructors had the option to record and post these lessons. However, legal complications arose: was there a need to obtain student consent to upload recordings of live classes in which they actively participated? With limited time to address every issue, most began with a tentative plan and adapted as circumstances evolved. Maintaining integrity of exams and assessments was a major concern for many and for the author in particular.

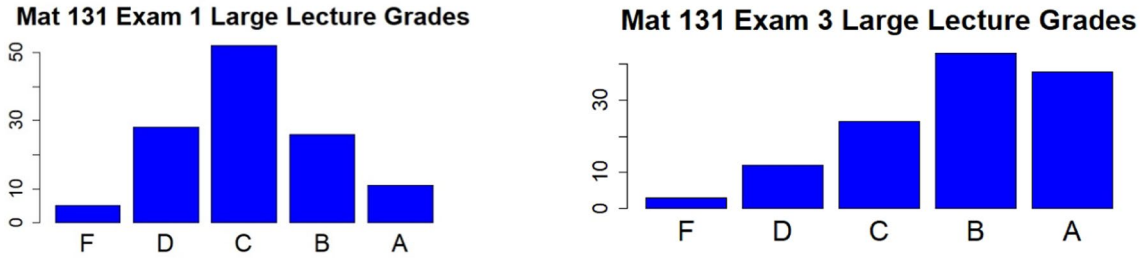

Fig. 2 The left distribution had been rather typical for all exams over the many semesters that the author has taught his large lecture statistics class, following a bell-shaped distribution. The distribution on the right comes from the author's very first online exam in Spring 2020. It has many more A's and B's and quite a bit less C's and D's

For the lower-level large lecture class, the author used the same format for exams as was used in the face-to-face exams: multiple choice questions. These were now given on the Blackboard system with no proctoring rather than in a proctored classroom setting. Whereas in the in-class setting, the author did not allow the use of book, notes, or cellphones (not to mention computers), there was really no way to prevent this in the online setting, so the author had to assume that students would be using all of these resources (despite whatever instructions were given). In addition, students could work with each other or get help from others in completing these exams. Exam collaboration with classmates can be made more difficult by scrambling both the question order as well as the order of the multiple choice answers for each individual exam —a feature available on Blackboard and other course management systems.

Blackboard, as well as CANVAS (another widely used system) has a feature called lockdown browser. This tool aims to prevent students from accessing other websites during their exams and prevents actions like copying and pasting. Although the author used this feature on his exams, there are ways around it—for example, by opening up a different browser. Using the lockdown browser also introduces potential complications. At times, it might malfunction, inadvertently preventing students from completing their exams—often through no fault of their own. While the malfunction rate isn't alarmingly high—in the author's experience, it hovers around 2–3%—in a class of 140 students, this equates to 3–5 individuals per exam in the large lecture that require special accommodation. Moreover, the necessity for the instructor to remain available during exams to address technical glitches and also to arrange new times for students to complete or retake their exams undermines one of the intended benefits of automated online exams. Implementing this feature to reduce cheating opportunities can introduce significant collateral challenges, both in terms of time and inconvenience.

Navigating through the initial semester of Covid, while simultaneously learning the intricacies of online teaching, proved to be a substantial challenge. Most of the author's energy was channeled into determining the most effective methods to deliver lessons and engage students. Foremost in his concerns was devising strategies to conduct exams with utmost integrity. It was evident that certain measures were essential to minimize cheating. One such measure was requiring that online exams were conducted within a narrow time slot. At first glance this may have seemed straightforward since all students should presumably be free during the designated time slot for the class. The unpredictable nature of technology meant that some students would inevitably encounter IT issues, necessitating buffer time to accommodate them. Within Blackboard, there are primarily two features to regulate the duration of an exam: one is the window during which the exam link is accessible, and the other is an inherent time limit on the exam itself.

With the first exam, I gave too much time and kept the exam difficulty level similar to that of an in-person exam. Both turned out to be major errors. The results are nicely illustrated by the following two bar plots of grade distributions for the students (in the same) class, in their first exam (that was administered in person), versus for their third exam (that was administered online); see Fig. 2 below.

NOTE: For readers outside of the United States: Grades in most US colleges and universities (as well as in primary and secondary schools) are letter grades which typically have the following meaning: A = outstanding, B = good, C = satisfactory, D = unsatisfactory, and F = failing. Grades are converted numerically with A = 4 points, B = 3 points, C = 2 points, D = 1 points, and F = 0. Using these numerical conversions, grade point averages for particular students (or groups thereof) and of exams can be computed.

Based on the author's impression of student understanding and participation during the online lectures as well as the extremely higher grades with the online exams, this discrepancy does not seem indicative of enhanced remote learning proficiency. Although this does not present a definitive proof that students were cheating with the online exams, the author feels that the elevated grades on Exam 3 can largely be attributed to the increased opportunities

for students to cheat during online assessments. There are other possible explanations, e.g., perhaps during the pandemic, students were able to study better at home with less distractions and reduced time spent on transportation. A controlled randomized experiment would reveal the true reason for the much higher grades on (the same) exams when given online, but, as we pointed out earlier, it would not be feasible to perform such an experiment.

Upon observing the results of Exam 3 during Spring 2020, it became apparent to the author that adjustments were necessary. A shorter exam duration was introduced, and the difficulty level was elevated. While these changes yielded a more "normal" grade distribution in Exam 4 (like the one on the left of Fig. 2), they risked unduly penalizing genuine students who refrained from any dishonest practices.

In the Fall 2020 semester some new trends became evident. Typically, the author would reuse questions on exams so as to compare student performances and gauge how certain modifications in teaching may impact student learning. The first three exams (out of a total of four plus a final exam) had similar grade distributions to the typical one (left of Fig. 2). This was attributed to the adjustments made by the author based on the learnings from the previous semester. This also gave a good indication that the usage of real-time tutors on exam assistance platforms was not drastically affecting the grade distribution. Perhaps most students (who used online learning platforms on exams) did not feel they needed a live tutor, relying instead the large question/answer banks that such platforms provide.

For the fourth exam, the author did a simple experiment: For Fall 2020 semester, the author gave exactly the same fourth exam as was used in a semester several years back when the course was given face-to-face. Sure enough, compared with the typical symmetric distribution (left of Fig. 2) with all pre-Covid exams, the grade distribution for the online exam was ridiculously skewed—A was now the most common grade, see Fig. 3.

The difference of the distribution in Fig. 3 with a typical symmetric distribution for the author's Exam 4 pre-covid were even more extreme if we take into account the exam curves. The reason is that all exams are typically curved (with very similar curves for each exam over all pre-Covid semester). But since the numerical scores were so much higher with the online exams, their curves were less generous.

For Exam 4, which tends to be the most difficult of all of the exams (including the final exam), the scores tend to be lower and so the curve is more generous. Here are the curves for Exam 4 both before and during Covid:

Pre-Covid curve for (face-to-face) Exam 4: A: 75–100, B: 63–74, C: 45–62, D: 30–44.

Covid curve for (online) Exam 4: A: 90–100, B: 80–89, C: 70–70, D: 50–69.

This shows that the differences in grades were even more extreme than the distributions indicated. For example, pre-covid on this exam, a 75% was the bottom of the A grade range, whereas for the online version a 90% was needed for the same bottom of the A. Indeed, if the covid curve for this exam were used to assign grades to the students who took the exam face-to-face, the distribution would be skewed to the right (see the right side of Fig. 4), with about half of the students getting an F.

For the online exam it would have been reasonable to make the curve even less generous: e.g. A: 95–100. But the instructor would never make exam curves stricter than the "standard curve" and before giving online exams there would have never been a need to do this, since the exams were always sufficiently challenging to necessitate some sort of a downward curve.

To counteract this, the author crafted a final exam using entirely new questions, aiming to level the playing field and diminish the advantages for those relying on external aids.

Moving on into the Spring 2021 semester, when the author taught the same large lecture statistics course online, additional experiments and modifications were introduced.

Two Experiments: The author crafted two distinct modifications for roughly half of the questions on the first exam:

(1) The modified question retained the original wording, and the multiple-choice answers remained unchanged. However, one numerical parameter (e.g., a standard deviation) was adjusted, altering the correct answer to another existing option.
(2) The original question's wording was preserved, but the initially correct multiple-choice answer was replaced. Along with this, a numerical parameter (e.g., a standard deviation) was adjusted to change the correct answer to a different option.

The results were quite illuminating. For Experiment (1) the average percentage of students who chose the previously correct (but now incorrect) answer for such questions was 72%! This is corroborating evidence that most students were using third party online platforms to cheat and that many did not even read the actual questions!

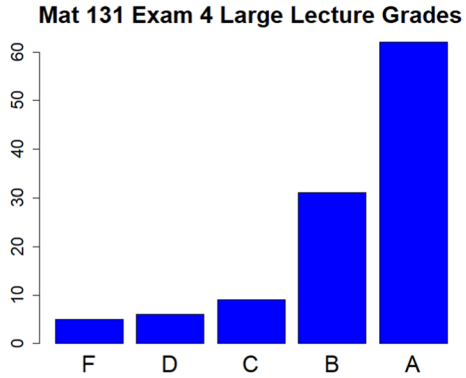

Fig. 3 The ridiculous distribution for Exam 4, of Fall 2020, which, as an experiment, used the same questions as a previous semester's exam a few years back. This provides compelling evidence that exam questions get added to third party exam assistance platform's data bases of exams and that most students were using such platforms to cheat

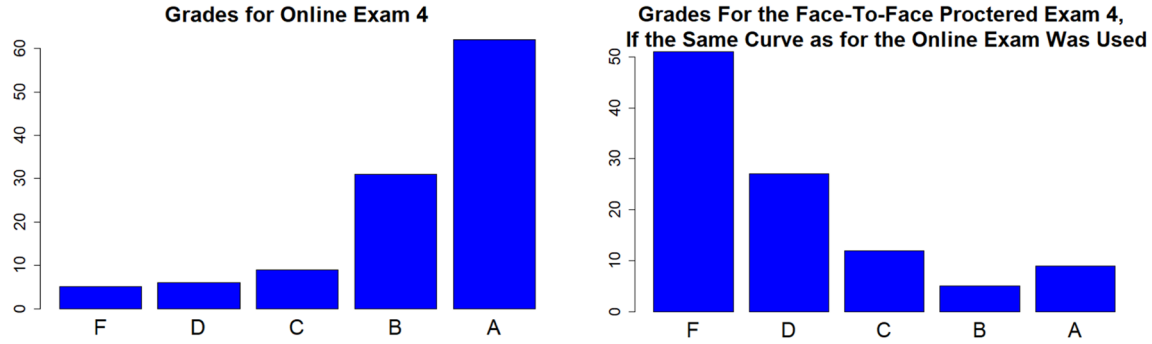

Fig. 4 A comparison of the grade distributions for the same multiple choice exam given online and unproctored (left) versus face-to-face proctored in the classroom (right) if the same curves were used

The 72% figure probably underrepresents the number of students who cheated in this way, as some astute learners might have read the questions and cross-referenced with their online platform to validate their initial interpretations.

In Experiment (2), students solely depending on third party online platforms were likely taken aback. Some even audaciously emailed the instructor, asserting that certain multiple-choice questions lacked the correct answers (which, indeed, were present).

4.2 Part 1B: introductory-level large lecture statistics courses after chat GPT

The emergence of Chat GPT has added further challenges for educators striving to uphold the integrity of online exams. As outlined in the introduction, Chat GPT can accurately answer numerous questions when they're directly inputted into the application. Educators are advised to test some of their online or take-home exam questions on Chat GPT to gauge the accuracy of its responses.

The capability of Chat GPT as well as other evolving large language models continues to improve. Questions for which Chat GPT gives the correct answer should be changed or replaced. We will give some general suggestions, but we first present the following rather revealing example.

Example

Original Draft Exam Question: Suppose that two dice are rolled and the numbers displayed on top are added up. The probability that this sum is at most 3 is…

The correct answer is: 1/12.

The Chat GPT answer is: 1/12 (= 3/36)—CHAT GPT Got the correct answer!!

Modified Draft Exam Question: Suppose that we have two pyramid shaped 4-sided dice, with the faces numbered 1 through 4. Suppose that the two dice are rolled and the numbers displayed on the bottom are added up. The probability that this sum is at most 3 is…

The correct answer is…3/16.

Chat GPT's Response: Incorrect.

Despite the current version of Chat GPT not answering the modified question accurately, its capability and efficiency continue to improve. It might be able to answer similar questions correctly in the future.

4.3 Timeline for the online classes above

Below is a timeline that summarizes the exam and grade information presented above:

- Spring 2020 Semester (Covid-19 started): Exams switched mid-semester from face-to-face to online. Exam 3 (the first online exam) had much higher grades than usual. Adjustments were made for remaining exams to mitigate this (shorter duration, increased difficulty).
- Fall 2020 Semester (first full online semester): The author continued with the above strategy as well as some new ones that were be elaborated upon in the following section (recommendations) for the first three exams and for the final exam. For Exam 4 an experiment was done. The same exam that was used several years ago was given. The grades turned out ridiculously high.
- Spring 2021 (another online semester for this class) The exams continued to be modified as explained above and this kept the exam results more down to earth.

Advice for making online exams for introductory statistics and related classes: The above discussion shows the variety of online resources that students can use to cheat on unsupervised online or take-home exams.

1. *No Repetition*: Third party online exam/homework assistance platforms have demonstrated that it's imperative not to reuse questions. Even those used just a day earlier can end up on such platforms. For integrity, either:
- Draft entirely new questions.
- Modify previous ones by adjusting parameters and/or altering the multiple-choice options.
2. *Tackling Live Tutoring:* It's a bit more challenging to address students who employ live tutors during exams. However, imposing strict time constraints can discourage this behavior.
3. *Chat GPT Screening:* As the efficiency of tools like Chat GPT increases, it's crucial for instructors to keep abreast of such new technology and to screen exam questions using a Chat GPT session to ensure they aren't easily answered by such platforms.
4. *Incorporating Graphics and Unique Symbols:* Questions that reference specific graphics or use specialized Greek letters in unique fonts can pose a challenge for platforms like Chat GPT. It adds an additional layer of difficulty in deciphering and interpreting the questions.

While these strategies involve more preparation, the integrity they bring to the examination process is invaluable. It's a trade-off: increased time in question creation versus the time saved from manual grading. Many educators, when given a choice, might still lean towards supervised, in-person examinations. However, for purely online courses, these precautions become essential.

5 Guidelines for online assessments in advanced mathematics and computer science courses

In general, upper-level STEM classes are not as susceptible to online cheating as the more standard lower-level classes. Many upper-level classes use specialized vocabulary and definitions that can vary even with the same course, depending on the instructor. Such vocabulary variations may also occur with introductory-level courses, but at a much smaller scale.

Upper-level math (post calculus) and computer science questions also appear in third party online exam/homework assistance platforms, but these are more difficult for any platform tutors to crack and keep up with.

1 *Specialized Vocabulary and Concepts*
- Upper-level courses frequently employ specific terminology and concepts that might differ based on the instructor or the curriculum.
- Variations in vocabulary make it more challenging for third party online exam/homework assistance platforms to provide accurate solutions.

2. *Computational Questions*
- While engines like Chat GPT and Wolfram Alpha can solve certain computational questions, exam setters should ensure that the questions are crafted uniquely to prevent direct solutions from these platforms.

3. *Proof-based Questions*

Questions that require proofs, common in many advanced math courses, and some computer science subjects, are generally tougher for both third party online exam/homework assistance platforms and Chat GPT to handle.

4. *Coding Questions*

Chat GPT has the ability to generate computer program codes. For questions requiring a computer code/program: Programs must run and produce an output and if the instructor tests the program, it should produce the same output. The instructor requires that students provide their codes in a format that can be copied and pasted. It is also required that the codes use only functions and protocols that were introduced in class. It is allowed to use new functions and constructs, but anytime such a usage is done, the new function/construct needs to be thoroughly explained with examples. Students are also informed that if their codes appear to be beyond the scope of the class's programming skills, the instructor may call them in after the exam to fully explain how their code works. Prior to Chat GPT, the author would allow students to use any new code strategies that were not taught in class without restriction; indeed, students are encouraged to go beyond expectations and study more programming methods. What usually happens when Chat GPT writes a computer program (in whatever programming language you ask for), it will feel free to use ANY functions and methods that it sees fit. Sometimes the functions used require the installation of some additional packages in order to run. This makes it fairly easy to detect when a student has used Chat GPT. Most often the programs will not run at all (in the author's machine learning course the software R was used) and/or, the program will use some constructs that are more advanced or simply different than were taught in class. Thus it should be required that the programs the student actually provide will indeed be executable (and they should provide the resulting output). If you were to ask a student who used Chat GPT to write their code to explain their code in case more advanced constructs are used, they will most often become flummoxed. In order to avoid the necessity of such encounters to verify whether the computer code was indeed the student's creation, it is simpler to simply not accept solutions that violate either of the above directions (or perhaps say that such solutions can earn no more than, say, 25% of the point values) in such cases.

5. *Course-Specific Assessments*

In other upper-level courses such as Cryptography, Discrete Mathematics, and Graph Theory, it is vital to frame questions based on the specific teachings of the course. This not only tests students' understanding but also mitigates the impact of online test support services.

In conclusion, while advanced courses present their own set of challenges for online assessments, a keen focus on course-specific teachings and a proactive approach towards leveraging technology can help ensure exam integrity and fairness.

6. *Conclusion*

The rapid transformation of the educational sector, catalyzed by unforeseen circumstances such as the pandemic, underscores the importance of adaptability and forethought. With distance learning not merely being an option but a necessity, online assessment, once considered a subsidiary component of education, has gained paramount importance. Our exploration unveils a multifaceted reality: the sheer potential of online learning, the unprecedented challenges

posed by third party online exam/homework assistance platforms and ChatGPT, and the undeterred perseverance of educators in ensuring academic integrity.

Our findings highlight an essential paradox. The very tools, like AI and internet platforms, that hold the promise to revolutionize and democratize education, also present profound challenges to academic integrity. As technology continues its relentless march forward, the academic community needs to be one step ahead, always innovating and strategizing to maintain the sanctity of education.

While third party online exam/homework assistance platforms exploit the vulnerabilities of the current online assessment system for profit, AI models like ChatGPT offer a glimpse into the future of information accessibility. They underscore the urgency with which educators and institutions must rethink and reshape assessment methodologies. The traditional paradigms of assessments, predominantly reliant on rote memory and standard problem-solving, may increasingly become obsolete. Instead, a greater shift towards analytical thinking, application, and synthesis may not only align better with real-world challenges but also be less susceptible to the shortcuts technology offers.

From introductory courses to advanced levels, the narrative remains consistent: there is no single foolproof method, no silver bullet. Instead, the solution lies in a combination of approaches—continuous adaptation, employing course-specific nuances, leveraging technology judiciously, and fostering an academic culture grounded in integrity.

According to the author's experience, the most important aspect for an online class to improve is the participation and peer communication of the students. I have noticed students have a much higher reluctance to participate in an online class than in an in-person class. There are many plausible reasons for this, for example, knowing that the class will be recorded and posted online might make students more hesitant to speak up. I am working on better understanding and mitigating this problem.

As we forge ahead into an increasingly digital future, it is imperative to view these challenges not as insurmountable barriers but as catalysts, driving us towards a more resilient, adaptive, and effective educational paradigm. It serves as a stark reminder that in the evolving dance between technology and education, staying static is not an option. The future of education demands foresight, adaptability, and a relentless commitment to academic excellence and integrity.

7. *Epilogue*

Readers might be curious about how the author's department and university are dealing with online courses with the learning experiences of the pandemic behind us. The author's campus belongs to the California State University system, which comprises of 23 campuses throughout the state and is the largest university system in the United States. While particular policies vary by department and campus, the system's higher administration, however, has made clear its preference to go back mostly to face-to-face instruction. At our college, in order to be offered online, any course must be approved by the department and the college dean, and then a university-wide committee. This academic year, our mathematics department has only two courses that have been approved for online instruction, one of which is the machine learning course that the instructor is currently teaching. A notable exception is our introductory statistics class, the main online course that was analyzed in this paper. This is the department's largest course, with over 40 sections offered every semester. The author continues to teach the (only) large lecture for this class, but at present there are no online sections being offered. As discussed, online courses have many advantages for democratizing education, but at the same time, there are important assessment issues that need to be further examined, so this topic will be an important area of research in years to come.

Author contributions A.S. is the sole author.

Funding Apart from the author's university salary; there was no additional funding for this paper.

Data availability All data for this paper was included in the paper, either in raw form or when more suitable via a graphical summary.

Code availability Not applicable.

Declarations

Competing of interests The authors declare no competing interests.

References

1. Sleator RD. The evolution of eLearning-background, blends, and blackboard…. Sci Prog. 2010;93(3):319–34. https://doi.org/10.3184/003685010X12710124862922.
2. Bartolic SK, et al. A multi-institutional assessment of changes in higher education teaching and learning in the face of COVID-19. Educ Rev. 2022;74(3):517–33. https://doi.org/10.1080/00131911.2021.1955830.
3. Oxford (University) Learning College webpage. https://www.oxfordcollege.ac/news/history-of-distance-learning/. Accessed 15 July 2024.
4. Bernard RM, Abrami PC, Lou Y, Borokovski E, Wade A, Wozney L, Huang B. How does distance education compare with classroom instruction? A meta-analysis of the empirical literature. Rev Educ Res. 2004;74(3):379–439.
5. Horspool A, Lange C. Applying the scholarship of teaching and learning: student perceptions, behaviours and success online and face-to-face. Assess Eval High Educ. 2012;17(1):73–88.
6. Horizon 2020 TeSLA Project. https://tesla-project-eu.azurewebsites.net/. Accessed 19 Feb 2024.
7. Fidalgo P, Thormann J, Kulyk O, et al. Students' perceptions on distance education: a multinational study. Int J Educ Technol High Educ. 2020;17:18.
8. Gamage KAA, de Silva EK, Gunawardhana N. Online delivery and assessment during COVID-19: safeguarding academic integrity. Educ Sci. 2020;10:301. https://doi.org/10.3390/educsci10110301.
9. Newton P, Essex K. How common is cheating in online exams and did it increase during the COVID-19 pandemic? A systematic review. J Acad Eth. 2023. https://doi.org/10.1007/s10805-023-09485-5.
10. Noorbehbahani F, Mohammadi A, Aminazadeh M. A systematic review of research on cheating in online exams from 2010 to 2021. Educ Inf Technol. 2022;27:8413–60. https://doi.org/10.1007/s10639-022-10927-7.
11. Dendir S, Maxwell RS. Cheating in online courses: evidence from online proctoring. Computers Hum Behav Rep. 2020;2:100033.
12. Golden J, Kohlbeck M. Addressing cheating when using test bank questions in online classes. J Acc Educ. 2020;52:100671.
13 Holden OL, Norris ME, Kuhlmeier VA. Academic integrity in online assessment: a research review. Front Educ. 2021. https://doi.org/10.3389/feduc.2021.639814.
14. Jia J, He Y. The design, implementation and pilot application of an intelligent online proctoring system for online exams. Interact Technol Smart Educ. 2021. https://doi.org/10.1108/ITSE-12-2020-0246/full/html.
15. Nguyen JG, Keuseman KJ, Humston JJ. Minimize online cheating for online assessments during COVID-19 pandemic. J Chem Educ. 2020;97(9):3429–35. https://doi.org/10.1021/acs.jchemed.0c00790.
16. Jeon J, Lee S. Large language models in education: a focus on the complementary relationship between human teachers and ChatGPT. Educ Inf Technol. 2023. https://doi.org/10.1007/s10639-023-11834-1.