

# INTELLIGENT TECHNOLOGIES IN EDUCATION

## Framing the threat posed by GenAI for online summative assessments: Incorporating Discord and higher-order learning

Joshua M. Wood

*Assistant Professor of Teaching, University of California–Riverside, Department of Political Science;*

[joshuamw@ucr.edu](mailto:joshuamw@ucr.edu)

### Abstract

Researchers have found that GenAI places considerable pressure on the integrity of online summative assessments. My research strengthens this conclusion in two ways. I argue that the impact of GenAI for online assessments is as widespread as are already popular messaging applications such as Discord. I argue furthermore, on the basis of a study utilizing problem sets from my own PHIL 007 Introduction to Critical Thinking, that utilization of GenAI on online summative assessments is not necessarily mitigated by devising problem sets which test for higher-order learning. ChatGPT-4o performed swimmingly on problem sets that consisted almost entirely of challenging critical thinking exercises. While these two results may seem to entail a dire situation for online summative assessments, the way forward for higher education in online spaces will require reshaping assessment validity so that learning outcomes are not compromised by student usage of GenAI.

### Keywords

Online assessment; summative assessment; academic integrity; ChatGPT; GenAI; Discord; critical thinking; higher-order learning; validity

### Citation

Wood, J. (2025). Framing the threat posed by GenAI for online summative assessments: Incorporating Discord and higher-order learning, *Intelligent Technologies in Education*, Advanced Online Publication.

### Editors

Section: Educational Technology

Co-Editor-in-Chief: Dr Marios Kremantzis

Associate Editor: A/Prof Xue Zhou

### Publication

Received: 23 December 2024

Revised: 18 March 2025

Accepted: 23 July 2025

Online: 28 July 2025

### Copyright

© by the authors, in its year of first publication. This publication is an open access publication under the Creative Commons Attribution [CC BY-ND 4.0](https://creativecommons.org/licenses/by-nd/4.0/) license.

## Introduction

I am an educator who uses online assessments and whose classroom policies about student utilization of GenAI are still evolving. The ubiquity of GenAI has forced me to think carefully about whether online summative assessments can serve as a legitimate measure of student learning. GenAI – understood as a program which, on the basis of a large language model, provides algorithmic text responses to user questions through a chat interface – presents a variety of opportunities and challenges for higher education. I believe that as rich as the research is on these topics, we have not yet fully expressed the challenge GenAI presents for online summative assessments. Summative assessments are those assessments which measure, on the basis of independent and unassisted student work, what a student has learned or where they stand with respect to the learning outcomes of a course. The counterpart, formative assessments, are those assessments used “to support learning” (Taras 2008, p. 172). Formative assessments are assignments where students, perhaps using additional resources (among which may very well be GenAI), are still in the phase of absorbing course content. If students lean on GenAI in order to complete online summative assessments, then their performance may undermine the value that these assessments hold within higher education.

The present study intends to add to the growing body of literature on the relationship between GenAI and higher education in two ways. I argue that the threat posed by GenAI has been underestimated in the case of online summative assessments. My argument is that Discord, a messaging application popular among students, cuts through one of the most powerful opportunity barriers: GenAI’s subscription models. Discord allows students to anonymously share the benefits of GenAI with other students who may not have access; and, perhaps more importantly, this messaging application lets students *passively* benefit from others who, submitting their solutions to the Discord channel, may have used GenAI productively in completing an assignment problem. I then present new data about GenAI’s performance on online assessments from my PHIL 007 Introduction to Critical Thinking. This data is limited to a single GenAI program, namely, ChatGPT which, at the time of this study, was the GenAI program most often directly mentioned to me during informal conversations with students. The PHIL 007 assessments were designed with the specific intent of fostering higher-order learning outcomes such as reasoning and spotting logical relationships between distinct claims. ChatGPT’s scores on these assessments, derived with minimal prompting, demonstrates its capacity to perform very well on questions requiring careful reasoning and creative applications of course content. This result suggests that what once seemed to be a limitation for GenAI – its inability to provide meaningful answers on questions which test for higher-order learning – has been surpassed in some academic disciplines.

My project adds to the data which serves as the basis of the growing concern, expressed by some researchers on assessment in higher education, that the integrity of online summative assessments is under

serious threat in the era of GenAI. The worry is not that GenAI makes *all* forms of online assessment unreliable as a measure of student learning outcomes, since some online assessments, particularly formative assessments, can be completed productively in collaboration with GenAI. The worry is that summative assessments, those reserved for measuring the application of valuable analytical and creative skills independently of the influence of GenAI, may be undermined by the ubiquity and capacity of GenAI in remote learning contexts. This is a dim view of the matter, one that contrasts with the overwhelmingly positive attitudes toward GenAI reported by students and educators alike (Ng et al., 2025; Soleimani et al., 2025). However, I, presumably like many educators delivering remote courses, am not prepared to discard online summative assessments. Therefore, I try to balance the results of my argument against considerations of validity in designing assessments. It is with a gesture of hope that I affirm the view, advanced by some researchers, that online summative assessments should be designed in such a way that their validity is preserved even with the expectation that students will be using GenAI to assist them. This approach may very well serve to address the challenge posed by Discord in online spaces.

## **GenAI, higher education, and online summative assessments**

It is by now clear that integrating student use of GenAI in higher education can be fruitful (Lodge et al., 2023; Perkins et al., 2024). As some argue, institutions of higher education owe it to students to actively incorporate GenAI in pedagogy and assessment so as to prepare students for the reality of the professional world: “Employers – and indeed the wider world – will expect graduates to be able to engage with these tools” (QAA 2023, March, p. 1; cf. QAA 2023, May, p. 3). Hence there is a push among stakeholders not merely to accept student reliance on GenAI, but to actively incorporate it into the classroom such as to design assessments which encourage students to utilize GenAI. The thought here is that authentic assessments should bear a greater resemblance to the professional world, a world in which reliance on GenAI for ordinary operations is increasing and in which GenAI skills among employees are valued. For example, as Perkins et al. (2024) suggest, the most current model for an “authentic assessment” is one in which there is “a deeper integration of GenAI tools to prepare students for professional life” (5).

Be that as it may, proponents of GenAI usage among students still recognize that there are important limits. Yes, higher education should incorporate GenAI so as to prepare students for the modern professional environment, but, at the same time, our practices should avoid “deskill[ing] students” in other ways (QAA 2024, February, p. 2). Such a risk is not automatic, since, as we know, students can benefit from using GenAI. Rather this risk arises from less than constructive uses of GenAI or, we might say, from “very poor academic practice” (QAA 2023, May, p. 3). Note that these guidelines are provided by the

Quality Assurance Agency for Higher Education. The worry here is that some students, as a consequence of an overreliance or misuse of GenAI, may suffer a reduction in their ability to conduct “evidence-searching” (QAA 2024, February, p. 2); or may end up with “undeveloped critical reasoning skills” (QAA 2024, February, p. 11). Nikolopoulou (2024) notes that student utilization of GenAI introduces the risk of “reducing original thought and undermining critical thinking processes” (3; cf. Griffin et al. 2025; Suriano, et al. 2025; Bastani, et al. 2024). Relatedly, Open Universities Australia (2025) has developed a website to explain to students what to avoid when making use of GenAI. This website discourages any use of GenAI which interferes with a student’s ability to form “original opinions,” denies them opportunities to “build your own argument,” and impedes their ability to “do the critical thinking.” Something similar is offered by Deakin University (2025), which discourages students from using GenAI in such a way that denies them their “own pursuit of knowledge and creativity.” These guidelines from Open Universities Australia and Deakin University are linked on the website “Artificial Intelligence” which is developed by the Tertiary Education Quality and Standards Agency (2025). Imparting the capacity to think critically, among other forms of higher-order learning, is at the heart of higher education. It seems reasonable to believe that we need learning practices which instil such skills as well as assessments which measure whether students have gained them. This is why even those organizations, such Quality Assurance Agency (serving higher education institutions in the UK and abroad) and TEQSA (serving those in Australia), which promote the value of GenAI in higher education ultimately recommend a balanced approach. QAA itself emphasizes the importance of assessments which test for and “reward higher-level skills” but which, at the same time, “are less susceptible to influence by Generative Artificial Intelligence tools” (May 2023, p. 3).

A pedagogical practice which explicitly combines both goals is the Two-Lane Approach developed by Liu & Bridgeman (2023). These researchers recognize the challenge of balancing, on the one hand, assessments which impart professionally relevant GenAI skills to students and, on the other, assessments which foster higher-order learning outcomes. It is a balance, as they put it, between “human-AI collaboration” and the “assurance” that students have met expectations for “all learning outcomes.” The second lane is, as they describe it, “unsecured.” It marks an approach within which formative assessments are either intentionally designed with human-AI collaboration in mind, or which are designed under the assumption that students “may (and likely will)” utilize GenAI. By contrast, the first lane is an approach in which summative assessments bar student usage of GenAI and favours in-person summative assessments of various forms among which they list traditional invigilated exams. While summative assessments in this first lane are not necessarily immune to academic misconduct, they are feasible ways of reliably measuring what students have learned without the risk of GenAI artificially inflating their performance. Liu & Bridgeman (2023), offering a two-lane approach, represent one among many other possible frameworks for

designing assessments in light of the ubiquity of GenAI. Other frameworks branch out into more nuanced distinctions, moving beyond all or none dichotomy.

The Traffic Light System provides three categories (Corbin, et al. 2025b); the AI Assessment Scale provides five categories (Perkins, et al. 2024); and the University of New South Wales provides six categories (Steel 2024). Researchers have criticized these more sophisticated frameworks for various reasons. For example, Curtis (2025) has argued against the Two-Lane approach because it, offering only the options of *all* or *none*, grants too much leeway in its second lane. Competing frameworks have avoided this simplification, but, as Corbin, et al. (2025a) argue, once we move beyond the two-lane approach it becomes challenging for both students and educators to determine which GenAI functionalities fit into these more subtle classifications. Finally, taken as a whole, these frameworks, as Corbin, et al. (2025b) argue, ultimately depend on student compliance without any further assurances. On this point, Steel (2024), writing about the framework at UNSW, states: “Our assumption is that our students...won’t take impermissible shortcuts.” Be this as it may, these different frameworks for incorporating GenAI into higher education, like the “secure” first lane proposed by Liu & Bridgeman, include a category which completely forbids the use of GenAI. This is a category into which traditional summative assessments fall. And it is the relationship between online assessments and *this* category that presents a challenge. My focus in what follows concerns precisely this issue as it arises in the context of online assessments.

As appropriate as these frameworks are for the current state of higher education, one grappling with the prospect of an economy heavily impacted by GenAI, they represent pedagogical approaches which, when discouraging GenAI usage, are largely limited to in-person courses and so are not readily transferred to online learning contexts. The difficulty is that online courses do not have the benefit of “assured” assessments, that is, summative assessments conducted under direct supervision or in some way that can reliably prevent or discourage student reliance on GenAI. Liu & Bridgeman seem to recognize this specific challenge in online spaces when they write: “It is becoming increasingly clear that there isn’t going to be a way to design take-home assessments that are ‘AI-proof.’” The challenge is that online learning environments, regardless of the framework within which we moderate permitted student usage of GenAI, are deprived of the opportunity to offer summative assessments which abide by the strictest limits on using GenAI. Though, admittedly, not all summative assessments need be divorced from student usage of GenAI, moving forward my references to summative assessments will be to those forms of student engagement which *are* intended to be conducted by students without the assistance of GenAI.

## GenAI, online summative assessments, and academic integrity

We have seen above that formative online assessments can be completed by students assisted by GenAI without undermining assessment integrity. Matters are different for summative assessments which even proponents of student utilization of GenAI in higher education reserve as an assessment form to be completed independently of any reliance on GenAI. However, the literature on the relationship between online education and academic integrity does not always distinguish between formative and summative online assessments. The literature on this topic sometimes assumes that any online assessment is one in which students are to be barred from utilizing GenAI. My hope is that we can make use of this literature, which does not always distinguish between online assessments in which GenAI is a welcome resource and online assessments where usage of GenAI is to be discouraged, in order to draw a conclusion that is specific to online summative assessments. The conclusion emerging from this scholarship suggests rather strongly that the reliability of online summative assessments has been brought into doubt by the popularity and capacity of GenAI. While, generally speaking, both students and educators have been found to have positive attitudes toward GenAI, its adoption among educators is slower than that among students (Soleimani, et al. (2025). In what follows, I will focus on students and what researchers have learned about student usage of GenAI in online contexts.

Studies on this topic suggest that academic misconduct is considerable in online spaces. Dendir and Maxwell (2020) write: “Although the early literature was somewhat inconclusive on the extent of cheating in online courses, recently there is mounting evidence that it is substantial” (8). Recent studies have continued to support this conclusion (Alin et al., 2023; Henderson et al., 2023; Susnjak & McIntosh, 2024; Newton & Essex 2024). Among recent research on this topic, Newton (2024) tells us that students themselves “report it is easy and common to cheat in online exams” (273). There are a wide variety of ways to cheat in online settings. However, of especial interest lately is the relationship between a specific GenAI program, namely, ChatGPT, and higher education assessments. If students were to utilize it for purposes of inflating their performance on online assessments, would it help? Yes. Researchers have found that even earlier version of ChatGPT, such as ChatGPT-3.5, can pass numerous assessments at the college level (Choi et al., 2022; Chaudhry et al., 2023; Nikolic et al., 2023; Terwiesch 2023). A more recent version, ChatGPT-4, has demonstrated improved performance. As reported by Stribling et al. (2024), “we find that ChatGPT-4 performs comparably to an above-average or exceptional graduate student on examinations in the biomedical sciences” (4). Likewise, Chaudry et al. (2023) found that this version of ChatGPT “was able to perform at par, with the highest performing student(s) in the class” on many academic tasks (23). Cahill & McCabe (2024) consider the issue of whether ChatGPT is providing quality responses to prompts and queries from its users. They report, emphatically, “The answers are *good*” (1).

Their remark about ChatGPT, specifically, captures the attitude among researchers exploring the relationship between GenAI and college-level assessments. GenAI is currently performing exceptionally well on challenging academic tasks and could serve as a means for cheating students to misrepresent their actual capacities.

Some researchers, faced with these results about the performance of GenAI, have drawn the conclusion that online assessments are less useful than they once were (Kirshner, 2023; Susnjak & McIntosh, 2024). These researchers urge us to rethink how we utilize online assessments in a time during which GenAI is capable of passing them. When it comes to measuring student learning outcomes by means of summative online assessments, GenAI, in their eyes, poses a “risk to assessment integrity” (Nikolic et al., 2023, p. 581). This group of researchers still allow for the possibility that modifying the structure of online summative assessments will make them serviceable. The thought here is that if we design online assessments so that they require reasoning and critical thinking, then they will be immune or at least less susceptible to the influence of GenAI. This is an approach which aims at “prevention,” that is, at making student reliance on GenAI “redundant by deploying strategies that make academic misconduct harder to commit in the first place, while at the same time promoting learning and the assessment of higher order outcomes” (Newton 2024, p. 274). However, another group of researchers draws a much stronger conclusion. For them, online summative assessments have ceased to be a viable instrument with which to accurately measure student learning outcomes even in those cases where we have introduced assessments designs intended to render GenAI less than useful. As they see it, given the ability of GenAI to perform very well on difficult assessments, “the basic validity of online, unproctored summative exams is questionable, and they should be avoided” (Newton 2023, p. 1). Other researchers in this group have found that ChatGPT performs well even on assessments whose multimodality is specifically intended to thwart ChatGPT’s assistance. For example, the developers of the two-lane approach, Liu & Bridgeman (2023), suggest that there is no strategy for summative assessment design capable of deterring the use of GenAI, since GenAI was able to perform well on all fronts: “Assessments that are authentic, or personal, or contextualised, or multimodal, or localised – these can all be completed by new generations of generative AI.” Another study, developed by Newton & Xiromeriti (2024), incorporate this fact about GenAI performance and reach the following conclusion: “unproctored online examinations are no longer a meaningful summative assessment method” (793). In their view, GenAI is capable of meeting assessment standards that test for higher-order learning and so we cannot deter student usage of GenAI in online spaces through assessment design.

These above views about the future of online assessments are largely predictive. Those with the least optimistic view maintain, given the capabilities of GenAI, that online summative assessments are

currently without value or will soon be without value as tools for measuring student success. There is a yet another group of researchers who are less interested in whether ChatGPT *could* perform well on assessments at the college level and more interested in the issue of whether college students are *actually* using it for this purpose. There has been a longstanding need for data of this kind, and it is slowly trickling in. There are two studies that interest me. Chan & Ahn (2023) found that students are either not utilizing ChatGPT or are using it too ineffectively to improve their scores on online assessments. Kortemeyer & Bauer (2024) found that those who cheat on online assessments still favor known methods of cheating – for example, using a search engine or perhaps a website such as Chegg.com to find answers online. Their view is that students have yet to make the transition to ChatGPT which, in the eyes of students, remains a novelty. However, findings from these two studies may already be behind the curve, given the rapid changes in technology which are shaping student culture within higher education. ChatGPT and associated AI chatbots are merging with student academic culture and the capabilities of these applications are becoming increasingly familiar to students. Recent studies of student perceptions of ChatGPT reveal that students find ChatGPT to be a highly appealing resource. Many students find ChatGPT to be “easy” and “enjoyable” (Holland & Ciachir, 2024, p. 8); “fun and interesting” (Ifelebuegu et al., 2023, p. 338); “extremely convenient” and “easy to use” (Richards et al., 2024, p. 4); “enjoyable and entertaining” (Strzelecki, 2023, p. 10); and something with which, technologically speaking, they experienced “no issues” (Strzelecki, 2023, p. 10). It is not surprising, then, to learn, as Cahill and McCabe (2024) have recently found, that “AI usage is widespread among undergraduate students” (1; cf. Corbin, et al. 2024). Therefore, as the current literature suggests, we have evidence in favour of the view that a significant number of students are utilizing GenAI for various purposes. This does not necessarily imply that students are utilizing it for purposes of academic misconduct. But we can confidently state that all the ingredients for academic misconduct in the case of summative online assessments are in place.

## **The significance of Discord for online summative assessments**

Those who are pessimistic about summative online assessments should not be dismissed. While we may not have empirical confirmation of whether students are using GenAI in a way that undermines the integrity of online summative assessments, everything considered, the prospect of such behaviour seems imminent. It is true that GenAI is performing exceptionally well on college assessments. It is true that certain GenAI programs, such as ChatGPT, is an especially attractive resource to students. But there is a third pillar that has not yet entered the discussion. We should take the additional step of combining the two preceding factors, the capacity and popularity of GenAI, with the reality of messaging applications which students use for purposes of academic misconduct in the context of online summative assessments (Garg & Goel,



2022; Alin et al., 2023). There are numerous messaging services which students use for this purpose, but I will limit myself to Discord. Discord is a messaging application that students can use on a computer or on their phone. They can create chatrooms where other students in a course can gather anonymously to share answers as they complete an online assessment. As a form of cheating, this method has been around for several years. Costley (2019) found that, within the population of students surveyed, sharing answers with other students concurrently enrolled in an online course, sometimes gathering online in a designated chatroom, was “the most common way” for them to cheat (210). The authors report that this form of academic dishonesty was so ingrained in the student culture that it was done “without much thought” (210). This trend has continued to the present day. Malik et al. (2023) found that, of the students they interviewed, 63.1% of them were willing to reveal that they share answers to online examinations with friends “most of the time,” using mobile devices or text messages for this purpose (210). However, researchers have not yet explicitly considered the threat that GenAI poses in light of this already familiar, convenient, and popular method to facilitate collusion among students.

Once we combine these three factors – the capabilities of GenAI with respect to college assessments, the widespread use of GenAI among college students, and the prevalent utilization of Discord among college students to share answers from online assessments – then even those of us who remain optimistic about online education should begin to worry. It does not take a significant number of students using GenAI to pose a risk to online summative assessments. Rather it merely takes a handful of students, and perhaps as few as one or two, to undermine the integrity of online summative assessments for a given course. A small group of students can use GenAI to gain correct answers to a question, and they can then take the simple and, for some, routine step of sharing these answers anonymously with a collective of other students enrolled in the same course. Moreover, the existence of Discord cuts through the issue of unequal access to GenAI, that is, to the extent that unequal access may be thought to shape academic dishonesty. It is true that GenAI’s most capable versions require subscription fees and thereby shut out students who do not have the money to pay for it. This will mean that students denied the opportunity to engage with GenAI’s most advanced versions will develop fewer skills in designing prompts in comparison to students who can afford to access and so directly utilize these versions of GenAI for themselves (Holland & Ciachir 2024). Which is to say that GenAI’s subscription model will indeed increase the “digital divide” among students (Ifelebuegu et al. 2023, p. 341; QAA 2023, May, p. 5). This is, of course, a point of concern. However, the subscription model, given the prevalence of messaging applications like Discord significantly expands the point of impact for programs such as ChatGPT. All things considered, the influence of GenAI will not necessarily be restricted to those students who can afford, for example, a subscription to the most recent and most advanced model of ChatGPT. Rather the influence of GenAI on online summative

assessments will likely occur where Discord *already* exercises an influence on student performance in online settings.

## Research question

My PHIL 007 Introduction to Critical Thinking is a skills-based course offered across a ten-week quarter at the University of California–Riverside (UCR). It incorporates a number of assignment formats, including proctored in-person paper exams and regular pop quizzes during lecture. However, my focus will be on its incorporation of MCQ online assessments. Some were categorized as homework (HW) and others were categorized as Quizzes. Here are three sample questions all of whose answers are (a):

Week 3 quiz: Question 4

Suppose we have an argument of the following sort:

- i. If Petra's car is not parked in her driveway, then Petra is at work today.
- ii. Petra's car is not in the driveway.
- iii. Petra is scheduled to work today.
- iv. Therefore Petra is at work today.

Which of the following methods, taken all by itself, *cannot* help us find fault with the argument?

- (a) Calling Petra's place of work to ask whether she is scheduled to work next week.
- (b) Looking to make sure that Petra's car is, in fact, not in the driveway.
- (c) Checking to make sure that Petra is, in fact, scheduled to work today.
- (d) Asking Petra's roommate whether Petra has parked somewhere other than in her driveway when she was at home.

Week 4 homework: Question 3

The following is an invalid argument:

- i. If I both eat a cheeseburger and drink a milkshake, then I will feel wonderful.

- ii. If I only eat a cheeseburger, then I will not feel wonderful.
- iii. If I only drink a milkshake, then I will not feel wonderful.
- iv. I ate a cheeseburger.
- v. Therefore I will feel wonderful.

Which of the following answer options *incorrectly* identifies what can be changed in order to make the argument valid?

- (a) The author needs to replace premise (i) with a simpler premise: “I drank a milkshake.” Once we join this revised premise (i) with the current premise (iv), then the author will have affirmed both antecedents stated in premises (ii) and (iii) whose consequents match up with the conclusion.
- (b) The author of this argument needs to introduce an additional premise affirming the second of the two antecedents in premise (i). As it is, looking at premise (iv), the author has only affirmed one of the two antecedents of the conditional statement in premise (i).
- (c) The author of this argument needs to alter the conclusion (v) to reflect the consequent in premise (ii). As it is, looking at premise (iv), the author has affirmed the antecedent in premise (ii) and so needs to change conclusion (v) so that it matches up with this conditional statement: “Therefore I will not feel wonderful.”
- (d) The author of this argument needs to change premise (iv) to say “I drank a milkshake.” This change would mean affirming the antecedent in premise (iii). But the author would also need to change conclusion (v) to match up with this conditional statement: “Therefore I will not feel wonderful.”

Week 5 homework: Question 1

Suppose we have an argument that looks like the following:

- i. If someone needs help and it is too hard to help them, then you shouldn't help them.
- ii. Petra needs help.
- iii. It is too hard for Sam to help Petra.
- iv. Therefore Sam should not help Petra.

Which of the following is *not* a useful criticism?

- (a) Showing that it is possible for someone to not need help at the same time that you should nevertheless help them.
- (b) Showing that it is possible both for someone to need help and for it to be very hard to help this person at the same time that you should nevertheless help them.
- (c) Showing that it is not, in fact, too hard for Sam to help Petra.
- (d) Showing that if it is true that Sam should not help Petra because of how hard it is to help her, then it follows that Sam, who loves his grandmother very much and would do anything for her, must also not help his grandmother whenever it becomes too hard to assist her.

During a typical quarter, students completed 100 HW MCQs and 100 Quiz MCQs online. I have designed these questions to require careful thinking on the part of students. The fact that these are MCQs does not take away from the challenge they present for students newly introduced to the basic elements of critical thinking. As Wisenhunt et al. (2022) argue, “when multiple-choice questions are well constructed..., they are capable of both assessing and encouraging deeper levels of processing” (141). I take questions such as those examples offered above, which make up a significant portion of the PHIL 007 online assessments, to serve as an opportunity for students to demonstrate their capacity for thinking critically or for solving problems creatively. In my estimate, these MCQs satisfy the recommendations laid out in Newton (2024) for crafting MCQs which “can be used to assess higher order learning” (275). If I am correct in taking these questions to test for higher-order learning, then this implies that using these problem sets to measure ChatGPT performance may yield interesting results.

ChatGPT has had several versions since its release as ChatGPT-3 on 22 November 2022 (Rudolph et al., 2023; Stribling et al., 2024). On 14 March 2023 ChatGPT-3.5 was made freely available to all users, and was joined by ChatGPT-4, which was restricted to paid subscribers. ChatGPT-4 was made available with limited use to free users on 6 November 2023. 13 May 2024 saw the introduction of ChatGPT-4o, which was restricted to subscribers while, after this date, ChatGPT-3.5 remained available to free users and ChatGPT-4 remained available with limited use to free users. Limited use, in this case, means that users can submit no more than ten of queries over a period of four hours. Once that time has passed, then the option to submit queries with ChatGPT-4 returns.

Three iterations of my PHIL 007 Introduction to Critical Thinking – which took place in Spring 2023, Fall 2023, and Spring 2024 – all coincide with one or more versions of ChatGPT: ChatGPT-3.5,

ChatGPT-4, and ChatGPT-4o. UCR runs on a quarter system in which its courses are packed into a ten-week period; its academic year is divided into three quarters, namely, Fall, Winter, and Spring. The Spring 2023 iteration of PHIL 007 took place from 4 April until 9 June 2023. For the duration of Spring 2023, ChatGPT-3.5 was freely available to users; and ChatGPT-4 was available to paid subscribers. The Fall 2023 course took place from 2 October until 12 December 2023. Again, ChatGPT-3.5 was freely available to users throughout this period, and ChatGPT-4 was still available to paid subscribers. But ChatGPT-4, exactly halfway through the Fall 2023 quarter, was made available for limited use to free users. The Spring 2024 quarter is the most complicated in terms of its relationship to ChatGPT's development. As with the previous two quarters, ChatGPT-3.5 was freely available to all users, and ChatGPT-4 was available to paid subscribers. This time around, however, ChatGPT-4 was also available with limited use to free users all quarter. Furthermore, at the conclusion of the second week and so for eight weeks of the ten-week quarter, ChatGPT-4o, the most advanced version, became both available to paid subscribers and available with limited use to free users.

With this context in place, let us take up the research question motivating the present study: *If students, differentiating them by quarter, had optimally utilized ChatGPT to assist their performance across these three iterations of PHIL 007, then what scores would students have received?*

## **Method**

ChatGPT was tested with the minimum number of queries. It was always tested with a single prompt that included a request that ChatGPT provide the correct answer to “the following multiple-choice question,” the question text, and the four answer options. No additional prompts were used and no further variation of the text was inputted. ChatGPT was fed specific problem sets for a given quarter so as to ensure that it was being tested with the same questions and answer options that were identical to the problem sets students encountered when taking the course. Problem sets differed slightly from one quarter to the next, since, among other small changes, new questions were added, answer choices were modified, and some questions were revised for greater clarity and precision. Furthermore, specific ChatGPT versions were tested on problem sets depending on which versions of ChatGPT were available during a particular quarter. For example, Spring 2023 problem sets were submitted to ChatGPT-3.5 and ChatGPT-4, but not to ChatGPT-4o, since the more advanced version was not available until one year later, during Spring 2024. However, ChatGPT-4o, which was introduced during the Spring 2024 quarter, was tested on the eight weeks of assessments with which it overlapped. Scores for ChatGPT-4o were averaged along with the results of

feeding the first two weeks of assessments to ChatGPT-4. Therefore, the ChatGPT-4o average represents the most optimal usage of ChatGPT-4 and ChatGPT-4o across the Spring 2024 quarter.

## Results: ChatGPT performance on PHIL 007 online assessments

**Figure 1: ChatGPT performance results**

Quarter	Assessment	ChatGPT-3.5	ChatGPT-4	ChatGPT-4o
Spring 2023	HW	77	95	
Spring 2023	Quiz	74	91	
Fall 2023	HW	72	94	
Fall 2023	Quiz	75	90	
Spring 2024	HW	75	91.25	98.75
Spring 2024	Quiz	71	92	95

The numbers in Figure 1 present the percentage of problems that ChatGPT was able to answer correctly for a given quarter either with respect to the total HW or total Quiz assessments offered. We can see that the least advanced version, ChatGPT-3.5, did fairly well. It received passing scores ranging from C- to C+ on both HW and Quiz assessments across this two-year period. ChatGPT-4 was tested on all the same materials as ChatGPT-3.5, and the former did much better, scoring in the A- to A range. The least ChatGPT-4 improved upon ChatGPT-3.5 was in Spring 2024 where the former still scored 16.25 points higher on the same assessments. ChatGPT-4o, the most advanced version of this AI application examined here, was available during (the last eight weeks of) a single quarter. Impressively, its scores are in the A to A+ range.

Let's look more closely at the feedback that ChatGPT gave in response to specific problems from the HW and Quiz assessments. We will take as our samples ChatGPT-4o responses to the three example problems offered above. I have provided headings so that we can differentiate the problems from one another. However, beyond this, I am providing the text, but not necessarily the formatting, exactly as it was given in response to the HW or Quiz problem.

Week 3 quiz: Question 4

The correct answer is: (a)

Calling Petra's place of work to ask whether she is scheduled to work next week. This method cannot help us find fault with the argument because it concerns a different time frame (next week) than the argument's timeframe (today). The other methods pertain directly to verifying the premises or the conclusion of the argument.

Week 4 homework: Question 3

The correct answer is: (a)

The author needs to replace premise (i) with a simpler premise: "I drank a milkshake." Once we join this revised premise (i) with the current premise (iv), then the author will have affirmed both antecedents stated in premises (ii) and (iii) whose consequents match up with the conclusion. This option incorrectly identifies what can be changed to make the argument valid because replacing premise (i) with "I drank a milkshake" does not lead to a valid conclusion that matches the given premises and conclusion. The other options correctly identify changes that align with the logic of the premises.

Week 5 homework: Question 1

The argument presented is: If someone needs help and it is too hard to help them, then you shouldn't help them. Petra needs help. It is too hard for Sam to help Petra. Therefore, Sam should not help Petra. To determine which of the provided criticisms is not useful, we need to analyze each option:

(a) Showing that it is possible for someone to not need help at the same time that you should nevertheless help them. This is not directly relevant because the argument concerns cases where help is needed, not cases where help is not needed.

(b) Showing that it is possible both for someone to need help and for it to be very hard to help this person at the same time that you should nevertheless help them. This directly attacks the main premise of the argument, challenging the idea that difficulty in helping someone automatically means you shouldn't help them.

(c) Showing that it is not, in fact, too hard for Sam to help Petra. This directly challenges the third premise of the argument. This directly challenges the third premise of the argument.

(d) Showing that if it is true that Sam should not help Petra because of how hard it is to help her, then it follows that Sam, who loves his grandmother very much and would do anything for her, must also not help his grandmother whenever it becomes too hard to assist her. This

provides a counter-example to the logic of the argument, suggesting that the principle might lead to undesirable conclusions in other similar situations.

Among these, option (a) is not useful as a criticism because it does not address the argument's premises or the logical structure directly. Instead, it brings up an irrelevant scenario that is outside the context of the given argument. Therefore, the correct answer is: (a) Showing that it is possible for someone to not need help at the same time that you should nevertheless help them.

## Discussion

Submitting multiple-choice question prompts to ChatGPT for the purposes of this study often yielded a lengthy step-by-step explanation that addressed what the question was asking, why certain answer options were unsatisfactory, and why one answer option was the most satisfactory of the bunch. Some of these were hallucinatory in the case of ChatGPT-3.5 where the reasons it provided for or against an answer option bordered on highly grammatical gibberish. But, in the case of ChatGPT-4 and ChatGPT-4o, the detailed explanations were, in my estimate, quite good. The explanations provided by ChatGPT were at least as good as any explanation students were likely to get from myself or a teaching assistant in office hours. Some explanations, considering their detail and thoroughness, might even be better. ChatGPT-4o's response to Week 5 homework: Question 1 is an example of what it sometimes offers. When a question and its associated answer options are complex, ChatGPT took the trouble of breaking it down so that the user could understand why it chose a specific answer option. Researchers have argued that, among its positive impacts for higher education, ChatGPT can provide "personalized instruction" for students (Rudolph et al. 2023; Ifelebuegu et al. 2023; and Choi et al. 2023). My findings, given the fullness of ChatGPT's explanations, suggest that this is indeed the case.

Moreover, my findings are consistent with Pardos et al. (2024). These researchers, using student performance as a metric, compared ChatGPT feedback on math problem sets with feedback offered by trained human tutors. What they found is that ChatGPT feedback yielded "statistically significant learning gains that were inseparable from human tutor-authored hints" (13). In the case of PHIL 007 problem sets, I was struck by how organized and meticulous the responses were from ChatGPT-4 and ChatGPT-4o. If students relied on these versions of ChatGPT to complete their problem sets *and* students read and thought about the explanations offered by ChatGPT, then, in my opinion, they would have learned a great deal about devising and applying strategies for thinking analytically. However, we can add a further point here about what ChatGPT offers to students. Instructors, tutors, and teaching assistants have limited availability shaped by their schedules and, even when available, are often tasked with the challenge of addressing the



needs of multiple students in less than adequate time. By contrast, ChatGPT provides fluid responses in seconds, responses that are specific to the prompts engineered by the user, and responses for which students can take the time to submit additional questions for purposes of further clarification. ChatGPT can do this at any time of the day and for as long as the user's ChatGPT account status permits. The issue of whether ChatGPT is capable of providing learning benefits to students is certain. It *can*.

When assessing the impact of GenAI on online summative assessments, some researchers have recommended that we design questions that test for higher-order learning. The assumption is that GenAI cannot provide satisfactory responses to such questions. Stokel-Walker (2022) discusses strategies for mitigating the utility which GenAI provides for assisting students with completing online assessments. This author suggests that “academics could respond by reworking written assessments to prioritize critical thinking or reasoning that ChatGPT can’t yet do.” Susnjak & McIntosh (2024), in their review of the current literature, report that there is a specific emphasis on averting the threat of GenAI by designing assessments “so that they require higher-order thinking and critical reasoning skills” (2). Other researchers and organizations have advocated for the same strategy (QAA 2023, May; Cochrane & Ryan 2023). I believe the problem sets for PHIL 007 Introduction to Critical Thinking would satisfy these researchers. The problem sets for PHIL 007 test for higher-order learning. The result of my study – most evident in the strong performance from ChatGPT-4 and ChatGPT-4o on HW and Quiz assessments seen in Figure 1 – suggests that this strategy for reducing the impact of GenAI on online summative assessments may no longer be sufficient. What I have found is that ChatGPT, even with minimal prompting, is able to easily answer questions designed with the intent of measuring higher-order learning. A student who is not content with the results provided by ChatGPT can direct this application to analyse the problem further. The present study, which, for purposes of efficiency, limited itself to submitting PHIL 007 HW and Quiz problems with a single prompt, does not mirror the ways in which students can engage with ChatGPT. Though a single prompt is all it took to generate very impressive results, students, in their interaction with ChatGPT, can ask multiple questions about specific aspects of a problem or about the distinct answer options.

If we combine the facility with which ChatGPT is able to successfully answer critical thinking questions, the ease with which students are able to get such results, and the fact that some students are willing to share these results with other students in a given course via messaging applications such as Discord, then it would seem to be the case that the threat level of GenAI for online summative assessments is catastrophic. Therefore, it is tempting to agree with the dim view mentioned above, one advanced by Newton & Xiromeriti (2024), according to which online summative assessments are no longer meaningful academic measures of student learning outcomes. This is not to deny that GenAI, in appropriate contexts, can provide educational benefits to students. But it does force institutions of higher education to confront

the fact that some online summative assessments may not offer reliable measures of whether students have made progress toward intended learning outcomes.

## **Limitations**

The present argument advances two claims. The first claim concerns messaging applications, such as Discord, which, through student collusion, can facilitate and amplify the impact of GenAI on online summative assessments. By ‘summative assessments’ I am referring to a special class of assessments – namely, those identified by researchers, institutions, and educational quality assurance agencies as deserving completion independently of student collaboration with GenAI. I believe that messaging applications do indeed extend the reach of GenAI. While this may be a point of concern for some educators, it is not necessarily an insuperable problem. The issue I want to raise here is not that there may be ways of effectively ensuring that students do not engage in sharing answers in this way. Rather the issue is that such collaboration may, in fact, yield beneficial learning outcomes. The significance of preparing students for a professional world in which collaboration is routine and expected may serve as a reason to permit or even encourage students to collaborate within the space of online summative assessments. Dawson et al. (2024) have argued persuasively on this issue that erecting barriers to student collaboration forces on them “an individualistic approach to work that is highly inauthentic” (1010; cf. Daumiller et al. 2023). Therefore, as we think about the impact of messaging applications on student utilization of GenAI in online spaces, we need to balance this against the important goal that institutions of higher education have with respect to preparing students for the modern workplace.

The second claim which I advance concerns the performance of ChatGPT on MCQs designed to test for higher-order learning. The purpose of the study is to add another datapoint to a growing body of literature which challenges the view that student utilization of GenAI can be discouraged, where such discouragement is warranted, through assessment design. I found that ChatGPT-4 and 4o performed excellently on MCQs designed to test for higher-order learning. Three limitations stand out here. The first is that it is indeed a single datapoint. Even if this datapoint is reliable, not all researchers need be convinced that GenAI is unstoppable in the area of assessment design. The second is that, while I am satisfied with the challenges the assessment questions presented to students, the MCQs on which ChatGPT’s performance was tested may not satisfy all researchers with respect to whether they effectively test for higher-order learning. The third limitation concerns the versions of ChatGPT which were tested. ChatGPT has already pressed ahead into more advanced models. Be that as it may, newer versions of ChatGPT would not prove

to be any more illuminating on this issue, since the older versions – ChatGPT-4 and 4o – already did quite well on these assessment questions.

However, there is a more pressing and much more interesting limitation which must be acknowledged. This is not a limitation of my specific ChatGPT study, but rather it is a limitation of a prevailing assumption on which this and similar studies are founded. The assumption is that online summative assessments cannot possibly stimulate the development of critical thinking skills unless students are discouraged from collaborating with GenAI. If this is our assumption – and it is an assumption for many educators, institutions of higher education, and educational quality assurance agencies – then the result of the present study, once combined with similar research results, do not look good for online summative assessments. However, as some argue, it is time to abandon or at least challenge this fundamental assumption about GenAI and student learning outcomes. I must refer again to Dawson et al. (2024) who develop a novel validity criterion to assess whether assessments, including online summative assessments, are appropriately designed. They write: “The use of artificial intelligence in assessment is not ‘cheating,’ it is a condition to be attended alongside other validity matters” (1012). Though Corbin, et al. (2025b) take a different path in their analysis of the current situation in higher education, they come to a very similar conclusion: “The goal is not to control AI use through rules but to design assessments that are not invalidated by the AI that is likely to be available to students when completing them” (p. 7). Hence, as Dawson et al. (2024) maintain, “assessments that depend on students not using artificial intelligence but are incapable of preventing students from doing so, are not particularly useful for high-stakes assessment of learning” (1012). Therefore, either online summative assessments, as traditionally conceived, should be recognized as unserviceable or we may need to revise the assumption that using GenAI *at all* is incompatible with developing critical thinking skills. Indeed, this assumption is weakening as researchers look more closely at the consequences GenAI has for students who rely on it. For example, as Essien et al. (2024) find, “ChatGPT enables students to engage more deeply with complex scenarios, thereby enhancing their analytical and evaluative skills” (876). While GenAI should not replace student efforts in these areas, these authors recognize that GenAI has “the potential...to foster a cognitive environment that effectively nurtures both fundamental and higher-order skills” (877). Likewise, the study from Pang et al. (2025) found that the “powerful capabilities of AI models such as ChatGPT can, if harnessed correctly, amplify student learning outcomes, particularly in critical thinking” (30). Similar results have been reached by Youssef et al. (2024); Bastani et al. (2024); Lee et al. (2025); and Wu et al. (2025). The challenge remains, however, with respect to *restricting* student usage of GenAI in online spaces to interactions which are productive for learning rather than merely passive, a challenge which, again, forces us to rethink assessment design. While more research is needed in this area, these studies push us to accept the possibility that one day online summative assessments may be conducted productively in full collaboration with GenAI.

## **Statement of interest**

The author attests that there is no conflict of interest associated with the results of this study or the argument developed in the space of the manuscript.

## **Funding statement**

The research conducted for purposes of this manuscript was supported by a 2024 UCR Regents Faculty Development Award and by UCR's Minority Serving Institution Research Academy, which operates within the Department of Political Science and which compensated interested RAs to serve as mentees and assistants for this research project.

## **Artificial intelligence use or non-use statement**

No artificial intelligence was used in connection with the creation of any element associated with this research project or manuscript.

## **CRedit author statement**

Joshua M. Wood: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing.

Allison Wang – Investigation

Jasmine Esmail – Investigation

Sarah Ochoa – Investigation

## **References**

Alin, P., Arendt, A., & Gurell, S. (2023). Addressing cheating in virtual proctored examinations: toward a framework of relevant mitigations strategies. *Assessments & Evaluation in Higher Education* 48(3), 262-75. <https://doi.org/10.1080/02602938.2022.2075317>

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, O., & Mariman, R. (2024). Generative AI can harm learning. SSRN, 4895486. <https://doi.org/10.2139/ssrn.4895486>

Cahill, C., & McCabe, K. (2024). Context Matters: Understanding Student Usage, Skills, and Attitudes Toward AI to Inform Classroom Policies. *PS: Political Science & Politics*. Published online 2024, 1-8. <https://doi.org/10.1017/S1049096524000155>

Chan, J.C.K., & Ahn, D. (2023). Unproctored online exams provide meaningful assessment of student learning. *PNAS* 120(31). e2303030120. <https://doi.org/10.1073/pnas.2302020120>

Chaudhry, I.S., Sarwary, S.A.M., Refae, G.A.E., & Chabchoub, H. (2023). Time to Revisit Existing Student's Performance: Evaluation Approach in Higher Education Sector in a New Era of ChatGPT – a Cast Study. *Cogent Education* 10, 2210461. <https://doi.org/10.1080/2331186X.2023.2210461>

Choi, E.P.H., Lee, J.J., Ho, M-H., Kwok, J.Y.Y., Lok, K.Y.W. (2023). Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education. *Nurse Education Today* 125, 105796. <https://doi.org/10.1016/j.nedt.2023.105796>

Choi, J., Hickman, K., Monahan, A., & Schwarcz, D. (2022). ChatGPT Goes to Law School. *Journal of Legal Education* 71(3), 387-400.

Cochrane, T. & Ryan, T. (2023.) ChatGPT and academic integrity: Options for adapting assessment in semester 1. Melbourne Centre for the Study of Higher Education. [https://melbourne-cshe.unimelb.edu.au/\\_data/assets/pdf\\_file/0008/4533218/ChatGPT-and-Academic-Integrity.pdf](https://melbourne-cshe.unimelb.edu.au/_data/assets/pdf_file/0008/4533218/ChatGPT-and-Academic-Integrity.pdf)

Corbin, T., Y. Liang, M. Bearman, T. Fawns, G. Flenady, P. Formosa, L. McKnight, J. Reynolds, & J. Walton. (2024). Reading at University in the Time of GenAI. *Learning Letters* 3: 35. <https://doi.org/10.59453/ll.v3.35>

Corbin, Thomas, Dawson, Philip, Nicola-Richmond, Kelli, & Partridge, Helen. (2025a). 'Where's the line? It's an absurd line': towards a framework for acceptable uses of AI in assessment. *Assessment & Evaluation in Higher Education* 1-13. <https://doi.org/10.1080/02602938.2025.2456207>

Corbin, Thomas, Dawson, Philip, & Liu, Danny. (2025b). Talk is cheap: why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education* 1-11. <https://doi.org/10.1080/02602938.2025.2503964>

Costley, J. (2019). Student Perceptions of Academic Dishonesty at a Cyber-University in South Korea. *Journal of Academic Ethics* 17, 205-17. <https://doi.org/10.1007/s10805-018-9318-1>

Curtis, Guy J. (2025). The two-lane road to hell is paved with good intentions: why an all-or-none approach to generative AI, integrity, and assessment is insupportable. *Higher Education Research & Development* 1-8. <https://doi.org/10.1080/07294360.2025.2476516>

Daumiller, M., Fritz, T., González Cruz, H., C. Rudert, S., & Janke, S. (2023). Cheating as a prosocial act? Helping others with academic cheating is related to social goals and cooperative norms. *Assessment & Evaluation in Higher Education*, 49(7), 948–961. <https://doi.org/10.1080/02602938.2023.2290979>

Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education* 49(7): 1005-1016. <https://doi.org/10.1080/02602938.2024.2386662>

Deakin University. (2025). Using generative AI. <https://deakin.libguides.com/generative-AI>

Dendir, S., & Maxwell, R.S. (2020). Cheating in online courses: Evidence form online proctoring. *Computers in Human Behavior Reports* 2, 100033. <https://doi.org/10.1016/j.chbr.2020.100033>

Essien, A., Bukoye, O. T., O’Dea, X., & Kremantzis, M. (2024). The influence of AI text generators on critical thinking skills in UK business schools. *Studies in Higher Education*, 49(5), 865–882. <https://doi.org/10.1080/03075079.2024.2316881>

Garg, M., & Goel, A. (2022). A systematic literature review on online assessment security: Current challenges and integrity strategies. *Computers & Security* 113, 1-13, 102544. <https://doi.org/10.1016/j.cose.2021.102544>

Griffin, H., & James, T. (2025). The temptation of the silver platter: Preventing the sidestep and normalising the productive struggle. *Intelligent Technologies in Education*, Advanced Online Publication. <https://doi.org/10.70770/tbb3db47>

Henderson, M., Chung, J., Awdry, R., Ashford, C., Bryant, M., Mundy, M., & Ryan, K. (2023). The temptation to cheat in online exams: moving beyond the binary discourse of cheating and not cheating. *International Journal for Educational Integrity* 19(21). <https://doi.org/10.1007/s40979-023-00143-2>

Holland, A., & Ciachir, C. (2024). A qualitative study of students' lived experience and perception of using ChatGPT: immediacy, equity and integrity. *Interactive Learning Environments*, 1-12. <https://doi.org/10.1080/10494820.2024.2350655>

Ifelebuegu, A.O., Kulume, P., & Cherukut, P. Chatbots and AI in Education (AIEd) tools: The good, the bad, and the ugly. *Journal of Applied Learning & Teaching* 6(2), 332-45. <https://doi.org/10.37074/jalt.2023.6.2.29>

Kirshner, S., Ryan, C., Toorn, C.V., Lim, E., & Cahalane, M. (2023, March 25). Education in the age of ChatGPT. <https://www.openforum.com.au/education-in-the-age-of-chatgpt/>

Kortemeyer, G., & Bauer, W. (2024). Cheat sites and artificial intelligence usage in online introductory physics courses: what is the extent and what effect does it have on assessments? *Phys. Rev. Phys. Educ. Res.* 20, 010145. <https://doi.org/10.1103/PhysRevPhysEducRes.20.010145>

Lee, Hao-Ping (Hank), Sarkar, Advait, Tankelevitch, Lev, Drosos, Ian, Rintel, Sean, Banks, Richard, & Wilson, Nicholas. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* 1121: 1-22. <https://doi.org/10.1145/3706598.3713778>

Liu, D., & Bridgeman, A. (2023). What to do about assessments if we can't out-design or out-run AI? <https://educational-innovation.sydney.edu.au/teaching@sydney/what-to-do-about-assessments-if-we-cant-out-design-or-out-run-ai/>

Lodge, J., Howard, S., Bearman, M., & Dawson, P. (2023). Assessment reform for the age of artificial intelligence, TEQSA, November 2023. <https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/assessment-reform-age-artificial-intelligence>

Malik, A.A., Hassan, M., Rizwan, M., Mushtaque, I., Lak, T.A., & Hussain, M. (2023). Impact of Academic Cheating and Perceived Online Learning Effectiveness on Academic Performance During the COVID-19 among Pakistani Students. *Frontiers in Psychology* 14, 1124095. <https://doi.org/10.3389/fpsyg.2023.1124095>

Newton, P.M. (2023). The validity of unproctored online exams is undermined by cheating. *PNAS* 120(41), e2312978120. <https://doi.org/10.1073/pnas.2312978120>

Newton, P.M. (2024). Guidelines for Creating Online MCQ-Based Exams to Evaluate Higher Order Learning and Reduce Academic Misconduct, S. E. Eaton, ed., *Second Handbook of Academic Integrity*, Springer International Handbooks of Education, pp. 269-85.

Newton, P.M., & Essex, K. (2024). How Common is Cheating in Online Exams and did it Increase During the COVID-19 Pandemic? A Systematic Review. *Journal of Academic Ethics* 22: 323-324. <https://doi.org/10.1007/s10805-023-09485-5>

Newton, P.M., & Xiromeriti, M. (2024). ChatGPT performance on multiple choice question examinations in higher education: A pragmatic scoping review. *Assessment & Evaluation in Higher Education* 49(6), 781-98. <https://doi.org/10.1080/02602938.2023.2299059>

Ng, Joey, Tong, Mavis, Tsang, Eva Y.M., Chu, Kevin, & Tsang, William. (2025). Exploring Students' Perceptions and Satisfaction of Using GenAI-ChatGPT Tools for Learning in Higher Education: A Mixed Methods Study. *SN Computer Science* 6: 476. 1-17. <https://doi.org/10.1007/s42979-025-04010-4>

Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., ... Sandison, C. (2023). ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education*, 48(4), 559–614. <https://doi.org/10.1080/03043797.2023.2213169>

Nikolopoulou, K. (2024). ChatGPT for educational-academic activities: Preschool student teachers' experiences. *Intelligent Technologies in Education*, Advanced Online Publication. <https://doi.org/10.70770/r5aett21>

Open Universities Australia. (2025). “How you should—and shouldn’t—use ChatGPT as a student.” <https://www.open.edu.au/advice/insights/ethical-way-to-use-chatgpt-as-a-student>

Pang, T. Y., Kootsookos, A., & Cheng, C. T. (2025). Artificial Intelligence in Higher Education Learning: Transferable Skills and Academic Integrity. *Intelligent Technologies in Education*, Advanced Online Publication. <https://doi.org/10.70770/vrz6qe26>

Pardos, Z.A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLoS ONE* 19(5), e304013. <https://doi.org/10.1371/journal.pone.0304013>



Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment. *Journal of University Teaching & Learning Practice* 21(06). <https://doi.org/10.53761/q3azde36>

QAA. (2023, March). QAA Academic Integrity Network Meeting.

[https://www.qaa.ac.uk/docs/qaa/members/academic-integrity-network-meeting-march-2023.pdf?sfvrsn=758ab81\\_6](https://www.qaa.ac.uk/docs/qaa/members/academic-integrity-network-meeting-march-2023.pdf?sfvrsn=758ab81_6)

QAA. (2023, May). Maintaining quality and standards in the ChatGPT era: QAA advice on the opportunities and challenges posed by Generative Artificial Intelligence.

<https://www.qaa.ac.uk/news-events/news/qaa-publishes-additional-advice-on-generative-artificial-intelligence-tools>

QAA. (2024, February). Quality Compass: Navigating the complexities of the artificial intelligence era in higher education. <https://www.qaa.ac.uk/docs/qaa/news/quality-compass-navigating-the-complexities-of-the-artificial-intelligence-era-in-higher-education.pdf>

Richards, M., Waugh, K., Slaymaker, M., Petre, M., Woodthorpe, J., & Gooch, D. Bob or Bot: Exploring ChatGPT's Answers to University Computer Science Assessment. (2024). *ACM Transactions on Computing Education* 24(1), 1-32. <https://doi.org/10.1145/3633287>

Rudolph, J., Tan, S., Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching* 6(1), 342-63. <https://doi.org/10.37074/jalt.2023.6.1.9>

Soleimani, S., Farrokhnia, M., van Dijk, A., & Noroozi, O. (2025). Educators' perceptions of generative AI: Investigating attitudes, barriers and learning needs in higher education. *Innovations in Education and Teaching International* 1-16. <https://doi.org/10.1080/14703297.2025.2530767>

Steel, Alex. (2024). 2 lanes or 6 lanes? It depends on what you are driving: Use of AI in Assessment. <https://www.education.unsw.edu.au/news-events/news/two-six-lanes-ai-assessment>

Stokel-Walker, C. (2022, December 9). AI bot ChatGPT writes smart essays – should professors worry? *Nature: News Explainer*. <https://doi.org/10.1038/d41586-022-04397-7>

Stribling, D., Xia, Y., Amer, M.K., Graim, K.S., Mulligan, C.J., & Renne, R. *Scientific Reports* 14, 5670. <https://doi.org/10.1038/s41598-024-55568-7>

Strzelecki, A. (2023). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments* 32(9), 5142-5155. <https://doi.org/10.1080/10494820.2023.2209881>

Suriano, R., Plebe, A., Acciai, A., & Fabio, R. A. (2025). Student interaction with ChatGPT can promote complex critical thinking skills. *Learning and Instruction*, 95, 102011. <https://doi.org/10.1016/j.learninstruc.2024.102011>

Susnjak, T., & McIntosh, R. (2024). ChatGPT: The End of Online Exam Integrity? *Education Sciences* (14), 1-20. <https://doi.org/10.3390/educsci14060656>

Taras, Maddalena. (2008). Summative and formative assessment: Perception and realities. *Active Learning in Higher Education* 9(2): 172-192. <https://doi.org/10.1177/1469787408091655>

Tertiary Education Quality and Standards Agency. (2025). Artificial Intelligence. <https://www.teqsa.gov.au/guides-resources/higher-education-good-practice-hub/artificial-intelligence>

Terwiesch, C. (2023). Would ChatGPT3 Get a Wharton MB? A Prediction Based on Its Performance in the Operation Management Course. *Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania*.

Wisenhunt, B.L., Cathey, C.L., Hudson, D.L. & Needy, L.M. (2022). Maximizing learning while minimizing cheating: New evidence and advice for online multiple-choice exams. *Scholarship of Teaching and Learning in Psychology* 8(2), 140-53. <https://doi.org/10.1037/stl0000242>

Wu, Fan, Dang, Yang, & Li, Manli. (2025). A Systematic Review of Responses, Attitudes, and Utilization Behaviors on Generative AI for Teaching and Learning in Higher Education. *Behavioral Sciences* 15, 467. <https://doi.org/10.3390/bs15040467>

Youssef, Enaam, Medhat, Mervat, Abdellatif, Soumaya, & Malek, Mahra Al. (2024). Examining the effect of ChatGPT usage on students' academic learning and achievement: A survey-based study in Ajman, UAE. *Computers and Education: Artificial Intelligence* 7: 100316. <https://doi.org/10.1016/j.caeai.2024.100316>